THE GENERAL CORPUS OF THE MODERN MONGOLIAN LANGUAGE AND ITS STRUCTURAL-PROBABILISTIC MODEL[1]

Krylov S. A.

Institute of Oriental Studies, Russian Academy of Sciences, Moscow, Russia

The paper describes the General Corpus of the Modern Mongolian language (GCML), which contains 966 texts, 1 155 583 words. We also report a morphological analyzer for Modern Mongolian language (MML), a grammatical dictionary for 63 071 lexemes, a general table of morphological homonymy. The processor analyzes effectively 97 % of textual word forms which correspond to 76 % word forms from the inputs of the concordance to the GCML. MML can be described in its quantitative aspect, according to a structural-probabilistic model (SPM) of MML. SPM contains frequency dictionaries (FDs) of MML of different types: FDs of word forms, lexemes, grammatemes, root morphemes and allomorphemes, affixal morphemes and allomorphemes, flexionemes, grammemes.

SPM allows describing behavior of various language units in the written text from the quantitative point of view: their frequency, distribution in texts, compatibility with other units etc. It is possible to transform the usual structural model into a SPM, which is based on statistical analysis of texts (in this model units of language are considered as possessing "their weight", the language oppositions and relations are being measured).

The paper reports the top lists of some FDs: i. e. ranging FD of word forms (top-list of the upper 44 word forms having frequencies higher than 1700 ipm), ranging FD of lexemes (top-list of the upper 44 lexemes having frequencies higher than 2050 ipm) and ranging FD of grammatemes (top-list of the upper 44 grammatemes having frequencies higher than 2909 ipm).

Key words: corpus linguistics, modern Mongolian language, frequency dictionaries, quantitative approach in linguistics

[1] The present research has been done under the financial support of the Program of the Presidium of the Russian Academy of Sciences on corpus linguistics on 2011. (direction 4 — "Creating and development of corpus resources on the languages of the world". The corpus has been created together with doctor of philology G. Ts. Pyurbeev, candidate of philology Natalya S. Yakhontova and candidate of philology Maria P. Petrova, to whom the author of the present article expresses a cordial gratitude.

1. About the general corpus of Mongolian

The initial version of the corpus of the modern Mongolian language reported here includes the following text genres:

1) fiction texts of the XX century: novels and stories; sketches;
2) poetry of the XX century;
3) «The Secret Legend of Mongols» epos translated into the modern language;
4) selection of newspaper articles from the newspaper «Dajaar Mongol».

The corpus contains 966 texts (1 155 583 words length).

A morphological analyzer, a dictionary for 63 071 lexemes, a table of homonyms have been created. The corpus has been lemmatized and glossed (in the spirit of the Leipzig glossing rules[2]).

The morphological analyzer at present works within the StarLing environment[3]. At present the work is on the experimental stage, 97 % of text word forms (which correspond to 76 % of the word forms which are inputs in the concordance of word forms) can be effectively analyzed.

Today the total analysed (morphologically marked) graphic word forms (allolexes) in the corpus is 1 103 233. The total of graphic word forms (lexes) in the corpus is 1 155 583, i. e. 97 %.

In total the amount of different allolexemes in the vocabulary of the concordance to the corpus (and in vocabulary of allolexemes) is 89 190. The share of the recognized allolexemes is 67 531, i. e. 76 %.

Overall effectiveness of the morphological analyzer can be visually presented in table 0.

Table 0.

	In the corpus:	part in %	In the vocabulary:	part in %
total word forms:	1 155 583	100 %	89 190	100 %
analysed:	1 123 156	97 %	79 137	89 %
analysed lexically:	1 104 911	96 %	68 212	76 %
analysed grammatically:	1 121 478	97 %	78 456	88 %
analysed both lexically, and grammatically:	1 103 233	95 %	67 531	76 %

[2] See Lehmann 1982; Croft 2003; Kassevich 2011: 214–221; also www.eva.mpg.de/lingua/resources/glossing-rules.php. A case study of the use of such notation in Mongolic studies (concerning a corpus with a resolved homonymy) see, for example, Baranova an Say 2009: 10–16, 873.

[3] The StarLing software environment was created by Sergey A. Starostin (1953–2005), and later it's been maintained by Philipp S. Krylov.

2. The Mongolian language in quantitative aspect

On the materials of the General Corpus of the Mongolian language first attempts of describing the Mongolian language in its quantitative aspect have been done. The Structural-probabilistic model of the Mongolian language includes frequency dictionaries (FD) of the Mongolian language of different types: frequency dictionaries of word forms in the ranging and in the alphabetic order (direct and inversed), FD of the bases in ranging and in alphabetic order (direct and inversed), FD of inflections in ranging and in alphabetic order (direct and inversed), FD of grammatemes in ranging and in alphabetic order (direct and inversed), and in an ideographic order, FD of lexemes in ranging and in alphabetic order (direct and inversed), FD of flexionemes in ranging and in alphabetic order (direct and inversed), FD of affixal allomorphemes and of affixal morphemes in ranging and in alphabetic order (direct and inversed), FD of root allomorphemes and of root morphemes in ranging and in alphabetic order (direct and inversed), FD of grammemes in ranging, alphabetic, and in an ideographic order.

3. The Structural-probabilistic model of Mongolian

The problem of studying of the quantitative characteristics of the MML is urgent because the majority of these characteristics till now are unknown to the scientists in the absence of a representative and at least relatively balanced corpus of MML to provide the material to apply distributive-statistical methods allowing to make professionally high-quality FDs and quantitative grammars, describing rate of units of morphology, derivatology, syntax and lexicology.

The quantitative approach allows to classify texts according to language styles and genres in frameworks of which these texts were created. As the distinctions between these styles and genres «are mainly a statistical quality»[4] thus it is possible to base the statistical stylistics of the MML describing and classifying texts of the MML on the strictly objective basis.

The quantitative approach to texts opens a way to studying MML while the segments of texts which are objects of calculations, are correlated with the units of the MML. The Linguo-statistical method allows to describe behavior of different language units (letters, morphemes, words etc.) in written text from the quantitative point of view: frequency of the use of units, their distribution in texts of a different genre, compatibility with other units etc. “At the same time the generalized quantitative information on classes of units, on language designs (e. g., the data about average length of a word or a sentence, on the frequency of the use of grammatical forms in different syntactic functions etc.) is being accumulated. Such information deepens the description of units of language”[5]. For example, simple ascertaining of existence

---

[4] See Shaykevich 1990: 231.

[5] See Shaykevich 1990: 231.

of forms of plural of nouns in RL and ML is insufficient for revealing of typological distinctions if one does not consider quantitative distinctions in the text functioning of the corresponding units. «This makes it possible to transform the usual structural model of language into a structural-probabilistic model, which is based on statistical analysis of texts (in this model units of language are considered as possessing "their weight", the language oppositions and relations are being measured). The structural-probabilistic model of language is more realistic, it is especially effective in diachronic and typological studies (.)»[6].

4. Frequency dictionaries FDs of Mongolian

Below we provide examples of the top lists of some FDs. Numerical indicators in column C mean relative frequency (quantity of occurrences of the given unit per one million word forms[7]), in column D — quantity of texts in which the given unit occurs, in column E – the rank of the given unit.

4.1. Frequency of word forms in Mongolian

Tab. 1 provides frequencies of word forms. Column A — a word form; column B gives the approximate English translation of the word of ML (in its main meaning).

Table 1.

A	B	C	D	E
нь	his, her, its, their *[also def. art.]*	24463.32	666	1
гэж	that *[conjunction for the object clause]*	13884.12	478	2
юм	*[marker of rheme]*	10052.32	529	3
ч	the	9882.74	540	4
л	let	6798.44	441	5
энэ	this, he, she, it, they	6628.87	463	6
тэр	that, he, she, it, they	5801.78	421	7
нэг	one *[actantial]*	5658.16	399	8
би	I *[subject]*	5429.76	446	9
хүн	man, person	5260.18	493	10
байна	is, are, am	4850.10	437	11
шиг	like, similar	4619.10	526	12
хоёр	two, both, and	4590.55	409	13
минь	my *[also a def. art.]*	4161.43	496	14

[6] See Shaykevich 1990: 231.

[7] On the use of "ipm" unit (instances per million words) see Sharoff 2002; Sharoff and Lyashevskaya 2009: 9.

A	B	C	D	E
байгаа	be *[ger. imperf.]*	3666.56	340	15
бол	as to; *[marker of topic]*	3606.86	363	16
дээр	on	3339.53	461	17
чинь	your *[also a def. art.]*	3109.39	344	18
байсан	was, were *[ger. perf.]*	2916.46	337	19
их	big, very	2907.81	403	20
дээ	let it be	2858.49	293	21
юу	what *[nom.case]*; whether	2722.66	307	22
гээд	having said; having told	2625.77	275	23
чи	you *[subject.]*	2620.58	289	24
уу	whether	2414.67	321	25
бас	too; also	2359.30	349	26
билээ	is, are, am	2338.53	303	27
байх	be	2313.44	316	28
сайхан	nice, beautiful	2217.41	429	29
байлаа	was, were	2190.59	265	30
шүү	not so?; isn't it?; doesn't it? wasn't it?; weren't it? didn't it?; etc.	2188.86	264	31
гэсэн	has told; has said	2091.10	314	32
вэ	*[special question]*	2049.57	283	33
болж	becoming	2018.42	322	34
биш	not	1965.65	306	35
та	you *[subject.]*	1965.65	259	36
гэдэг	they say	1925.85	322	37
одоо	now	1884.32	283	38
миний	my, mine	1876.54	330	39
хар	black; look!	1816.84	315	40
газар	land; country	1812.51	331	41
үгүй	no	1795.21	282	42
хүний	of a man, of a person *[gen.case]*	1760.60	389	43
болсон	become	1700.91	317	44

4.2. Frequency of lexemes in Mongolian

Tab. 2 represents frequency of lexemes. In column A lexemes are given; in column B approximate English translations of the word of ML (in its major sense) are given.

As the work was done on the corpus with unresolved homonymy sometimes the status of lexemes is ascribed not actually to lexemes, rather to disjunctive bundles of (partially) homonymic lexemes. However the information on the frequency of such units in the corpus is not less important than the information on the frequency of the

real lexemes. Anyway, for the information revealing quantitative characteristics of lexicon and grammar of the ML it is not necessary to wait, when the corpus with the unresolved homonymy will be created: it is necessary to wait too long, and anyway the data received on the basis of the corpus with the resolved homonymy will be based on too small empirical facts that will depreciate their statistical significance.

Table 2.

A	B	C	D	E
нь	his, her, its, their *[also def.art.]*	24463.32	666	1
гэж	that *[conjunction for object clause]*	13887.58	478	2
байх	be	13638.41	525	3
юм~юм(ан)	*[marker of rheme]*; thing	10408.76	531	4
ч	the	9882.74	540	5
болох	become	9672.51	528	6
явах	go	7327.06	511	7
л	let it	6845.16	443	8
энэ	this; he; she; it	6636.66	463	9
тэр	that; he; she; it	6024.99	427	10
нэг(эн)	one	5749.00	402	11
хүн	person; man	5538.77	497	12
ирэх	come	5497.24	439	13
би	I	5434.95	448	14
байн~байх	be	5364.87	449	15
хоёр	two; both; and	5197.03	427	16
шиг~шигэх	like; similar	4624.29	526	17
дээр	on	4445.20	503	18
байг~бай~байх	be	4322.35	362	19
минь	my *[also def. art.]*	4163.16	497	20
хэлэх	speak; talk	3645.79	348	21
бол	as for *[marker of topic]*	3609.46	363	22
гарах	come out of	3329.14	392	23
чинь	your *[also def.art.]*	3110.26	344	24
орох	enter	3051.43	363	25
гэх	say; tell	3019.42	328	26
их	very; big	3011.63	411	27
дээ	let it	2863.69	294	28
юу~юу(н)	what? whether	2725.26	307	29
чи	you *[sg.]*	2661.24	307	30
бодох	think; count; calculate; intend; aspire	2637.88	303	31
гээд~гээ~гэх	having told; having said	2627.50	275	32
сэтгэл	thaought; intention; wish	2464.85	405	33
сайхан	beautiful; nice	2448.41	458	34

A	B	C	D	E
уу~уух	whether; drink!	2424.18	322	35
харах	look	2386.12	343	36
бас	too; also	2365.35	350	37
билээ	is; are; am	2340.26	303	38
мэдэх	know	2265.86	316	39
үзэх	see	2157.71	355	40
өгөх	give	2127.43	307	41
гэсэх~гэх	say; tell	2101.48	314	42
вэ	*[marker for special question]*	2052.16	284	43
гэдэг~гэх	they say	2050.43	325	44

A lot of useful information can be drawn based on top-lists of language units in ranging dictionaries. Such information gives impulse to some interesting observations of typological nature. Below we attempt to present some of such observations.

The conjunction *бөгөөд* 'and' occupies the 52nd position in the FD of wordforms and the 67th position on the FD of lexemes.

Compare: the conjunction *and* in English FDs occupies the 3rd (or 4th , or 5th) position; conjunction *u* ('and') in Russian FD occupies the 1st position.

Why such difference? The difference can be explained in the following way. Mongolian prefers asyndetic constructions. The frequent asyndeton usually is being compensated through the phenomena of the so-called Altaic type of coordination which means the group inflection of nouns and the rich system of gerundial taxis constructions in the domain of verb.

2. Possessive pronouns have extraordinary high frequencies.

E.g., *нь* (his, her, its, their) occupies the 1st position in the FD of lexemes and the 1st position in the FD of word-forms. Compare: in English FDs *his* occupies the 25th (or the 23rd, or the 12th) position, *her* occupies the 42nd (or the 29th, or the 13th) position, *its* occupies the 78th (or the 77th, or the 142nd) position, *their* occupies the 36th (or the 39th, or the 61st) position. In Russian FDs *его* ('his', 'its') occupies the 41st (or the 50th) position, *её* ('her') occupies the 72nd (or 121st) position, *их* ('their') occupies the 86th (or 134th) position.

минь ('my') occupies the 14th position in the FD of word-forms and the 20th position in the FD of lexemes. Compare: in English FDs *my* occupies the 44th (or the 34th, or the 24th) position. In Russian FDs *мой* ('my') occupies the 60th (or the 69th) position

чинь ('your') occupies the 18th position in the FD of word-forms and the 24th position in the FD of lexemes. Compare: in English FDs *your* occupies the 69th (or the 64th, or the 62nd) position. In Russian FDs *твой* ('your') occupies the 266th (or 579th) position

Why such difference? The difference can be explained in the following way. Mongolian has a grammatical category of possessiveness. It can be expressed synthetically or analytically. Analytic means of expression of the category are the encliticized possessive pronouns. In fact the encliticized possessive pronouns play a kind of role which is fulfilled by articles in the European (e. g., Romance and Germanic) languages with articles. If one compares the European possessive pronouns with the Russian ones,

one can see that the European possessive pronouns are used more frequently than the Russian ones (in Russian-English translations such facts can be seen especially clear: compare such translational equivalents as *жена* ó *my wife* (*your wife*, *his wife*), *мать* ó *my mother*(*your mother, his mother, her mother, our mother, their mother*), *муж* ó *my husband*(*your husband, her husband*) , *отец* ó *my father*(*your father, his father, her father, our father, their father*), *нос* ó *my nose*(*your nose, his nose, her nose*), *голова* ó *my head*(*your head, his head, her head*) etc."). Comparison of FDs demonstrate that Mongolian (like other Altaic languages) has moved further then European languages in the scale of grammaticalization of possessive pronouns. As an important result of this movement evolves a high degree of desemantization of of possessive pronouns: such pronouns play rather the role of definiteness marker, of topic marker or a substantivization marker, than the role of possessivity markers. But such process is own to all grammatical categories, so the fact of desemantization of the possessive markers prove its grammaticalized nature.

4.3. Frequency of grammatemes morphological tagsets in Mongolian

Tab. 3 represents frequencies of grammatemes morphological tagsets. Column A represents grammatemes; column B gives an explanation of the used notation.

As the study was done on the corpus with unresolved homonymy sometimes the status of morphological tagsets is ascribed not actually to grammatemes, rather to disjunctive bundles of (partially) homonymic grammatemes. Nevertheless the information on the frequency of such units in the corpus is not less important than the information on the frequency of real grammatemes. What was said above about the reason about effectiveness of use of the corpus with unresolved homonymy mutatis mutandis concerns also the grammar.

The fact that there are some disjunctive morphological tagsets in which the left member of a disjunction is identical with the right, can be explained by the fact that in ML there are many partially-homonymic pairs (a) in which the relation of subcategorial conversion (namely, one of members of pair belongs to thematic declination, and another — to athematic) takes place, and also (b) members of which differ in such a way that the final element of one of the members of such pair contains a steady «н», and the final element of the other member of the pair contains an unstable «н».

Table 3

A	B	C	D	E
NOM	nominative case	281693.26	885	1
0	the unique form of an uninflected word	93456.52	865	2
CVB.CNGR	gerund, congressive	50195.83	806	3
NOM~GEN		34055.37	857	4
PC.PROSP-NOM	participle, prospective	30514.26	794	5

A	B	C	D	E
CVB.MOD	gerund of manner	23820.50	733	6
NOM~CVB.MOD		23777.25	777	7
PC.PRF-NOM	participle, perfective, nom. case	23658.72	767	8
NOM~ABS-NOM		23602.48	753	9
GEN	genitive case	16387.03	779	10
NOM~VF.OPT.IMP		12322.50	675	11
DAT	dative case	11860.50	753	12
VF.IND.AOR	predicative role, indicative mood, aorist	11502.33	489	13
POSS.REFL	reflexive-possessive	10403.57	664	14
NOM~DAT		9257.23	639	15
REL-NOM	nominative case, attributive role	9122.27	662	16
NOM~REL-NOM		8674.98	677	17
NOM~COM-NOM		8229.42	646	18
NOM~VF.OPT.JUSS		8070.23	666	19
CVB.ANT	gerund, antecessive	8031.30	548	20
PC.PROSP-DAT	participle, prospective, dative case	7071.83	490	21
VF.IND.PRS1	predicative role, indicative mood, present № 1	6936.00	590	22
GEN/ACC	truncated form of genitive-accusative	6495.64	573	23
ACC	accusative case	6383.16	647	24
ABL	ablative case	4746.28	532	25
PC.US-NOM	participle, usual, nominative case	4708.21	477	26
0~VF.OPT.IMP		4624.29	526	27
A.COM-NOM~DAT		4610.45	552	28
NOM~ACC~VF.OPT.JUSS		4435.69	377	29
VF.OPT.IMP	predicative role, optative mood, imperative	4057.61	457	30
NOM~DAT~CVB.ANT		3978.88	329	31
NOM~CVB.CNGR		3920.05	401	32
NOM~CAR-NOM		3806.71	461	33
NOM~GEN~GEN		3646.66	478	34
NOM~PC.PRF-NOM		3352.50	534	35
VF.IND.PROF-EMPH	predicative role, indicative mood, profect, emphatic	3308.38	370	36
NOM~PC. PROSP-CAR-NOM		3278.10	390	37

A	B	C	D	E
NOM~A.COM-NOM~DAT		3273.77	509	38
COM-NOM	comitative case, nominative case	3156.98	415	39
NOM~CVB.ANT		3146.59	363	40
DAT~CVB.ANT		3114.58	390	41
NOM~CVB.TERM		3023.74	369	42
CVB.MOD~PC.PRF-NOM		2951.93	365	43
INSTR	instrumental case	2909.54	449	44

A lot of useful information can be drawn based on top-lists of grammatemes in ranging dictionaries. Such information gives impulse to some interesting observations of typological nature. Below we attempt to present some of such observations.

1. Gerundial forms of verbs (so-called converbs) occupy high positions in the FD of grammatemes: e. g., congressive converbs — 50195.83 ipm, modificative converbs — 23820.50 ipm, antecessive converbs — 8031.30 ipm.

In West European (Romanic and Germanic) languages (unlike Uralic and Altaic) converbs are used with lower frequency. Why such difference? The difference can be explained in the following way. Mongolian is characterized by the so-called Altaic type of coordination that means a rich system of gerundial taxis constructions in the domain of verb. Thus, it prefers the asyndetic coordination. Rare use of conjunctions is being compensated with a rich system of converbs.

2. Reflexive-possessive forms are used extraordinary frequently ("bare" reflexive-possessive forms — 10403.57 ipm; reflexive-possessive forms of the perfective participles — 2336.80 ipm; reflexive-possessive forms of the dative case of the prospective participles — 2001.12 ipm, etc.). The most familiar for us European languages do not have such form at all. The Russian does have the nearest translational equivalent of it - the reflexive-possessive pronoun *свой* (literally, 'his own', 'its own', 'her own', 'my own' etc., or 'of himself', 'of herself', 'of myself' etc.). But if we compare the relative frequency of *свой* with the relative frequency of the reflexive-possessive forms, we see that *свой* is used with a lower frequency. Its relative frequency is 3825.5 ipm.

Why such difference? The difference can be explained in the following way. The reflexive-possessive forms of Mongolian are used as one of the main devices of expressing co-reference. Languages with articles prefer other way of expressing co-reference (mainly, definite articles). Russian prefers the so-called "zero" forms of expressing co-reference. In fact they have rather prosodic (supra-segmental) than "zero" nature; but the prosodic devices are expressed in the written form of language very rarely.

References

1. *Baranova V. V, Saj S. S.* Ot sostavitelej (From the composers), in Issledovanija po grammatike kalmytskogo jazyka (Researches on the grammar of Kalmyk language). (= Acta linguistica petropolitana = Trudy instituta lingvisticheskih issledovanij (Works of Institute of linguistic researches), vol.V, pt. 2) SPb.: Nauka, 2009, p. 7–21.
2. *Kasevich V. B.* Vvedenie v jazykoznanie (Introduction to linguistics). M.-SPb.: Akademija, 2011, 230 pp.
3. *Shajkevich A. Ja.* Kolichestvennye metody v lingvistike (Quantitative approach in linguistics) In Lingvisticheskij èntsiklopedicheskij slovar' (Linguistic encyclopedic dictionary). M: Sovetskaja Èntsiklopedija (Soviet Encyclopedia), 1990, p. 231.
4. *Ljashevskaja O. N., Sharov S. A.,* Chastotnyj slovar' russkogo jazyka (na materialah Natsional'nogo korpusa russkogo jazyka). (Frequency dictionary of modern Russian (on materials of the National corpus of Russian)). M: Azbukovnik, 2009. (= http://dict.ruslang.ru/freq.php)
5. *Croft B.* Typology and universals. Cambridge, 2003.
6. *Lehmann Ch.* Directions for interlinear morphemic translations. In Folia linguistica, 1982, Vol. 16, p. 193–224.
7. *Sharoff, Serge,* Meaning as use: exploitation of aligned corpora for the contrastive study of lexical semantics. In Proc. of Language Resources and Evaluation Conference (LREC02). May, 2002, Las Palmas, Spain, 2002.