УДК 001.103:002

АЙРЫМ ЖАНРДАГЫ КЫРГЫЗ ТЕКСТ КОРПУСТАРЫНДА СӨЗ ФОРМАЛАРЫНЫН ОРТОЧО КАЙТАЛАНЫШЫ ЖӨНҮНДӨ.

Б. Шаршембаев, Э.Бекбоев
Кыргыз-Түрк "Манас" Университети
ernisto.bee@gmail.com

Верификация средней повторяемости словоформ в некоторых жанрах корпуса кыргызского текста для различных объемов выборок. Сопоставительно-типологические характеристики коэффициента повторяемости генетически родственных и неродственных языков.

Verification of the average frequency of occurrence of the forms in certain genres corps Kyrgyz text for different sample sizes. Comparative-typological characteristics of the ratio of the frequency of occurrence of genetically related and unrelated languages.

Дүйнөлүк лингвистикадагы тилдик кубулуштарды заманбап өңүттөн изилдөө багыты ар тараптан өркүндөп өсүүдө. Мындай кубулуштарды иликтөө үчүн лингвостатистикалык ыкмалардын жана компьютер технологиялардын зарылдыгы өзгөчө.

Бул максатты ишке ашыруу жана иликтөөгө алынган тексттердин ишенимдүүлүк даражасын арттыруу үчүн өйдөкүлөргө кошумча төмөнкү тексттер да компьютерге жүктөлдү:

I	Ак Башат журналы, 2004-2012 жж.., № 1-26
II	Жаңы Ала-Тоо журналы, 2011 ж., № 21-32
III	Жаңы Ала-Тоо журналы, 2012 ж., № 33-44
IV	Жаңы Ала-Тоо журналы, 2013 ж., № 45-56
V	Шоокум журналы, 2005-2013 жж..
VI	Төлөгөн Касымбеков. Сынган кылыч: Тарыхый роман.—Б.: Кыргызстан, 1998.
VII	Койчиев Арслан Капай уулу.Айта бар менин кебимди... www.bizdin.kg
VIII	Койчиев Арслан Капай уулу.Мисмилдирик (Бедел белиндеги каргыш). Б.:Бийиктик 2009
IX	Бөртө Чоно. Чыңгызхандын өмүрү жөнүндө тарыхый роман. Которгон А. Саспаев — Б.: «Сүрөт-Басма-Салону», 2003. — 224 б.
X	Мартин Иден. Роман. Которгон Ж. Султаналиев. Ф., «Кыргызстан» 1977 468. бет.
XI	Джек Лондон. Өмүр кызык. Которгон С.Ерматов. Кыргыз мамлекеттик басмасы. Фрунзе – 1960
XII	Токтоналиев Жапаркул . Хан Ормон: тарыхый роман: 1-китеп: Б.: ААК «Акыл» басмасы, 20002. 596 б.
XIII	Майн Рид. Башы жок чабендес: Роман: Мектеп жашындагы тестиер балдар үчүн / Которгон Сүйүнтбек Бектурсунов; Сүрөт ред. Б. Жайчыбеков. — 2-бас. — Ф.: Мектеп, 1987.-432 б.
XIV	Медербек Адылбек уулу. Улуттун жоголгон байлыгы. 1-бөлүм. Үркүн. Б.:2011

Кыргыз текстинин статистикалык структурасын жыштык сөздүктөр аркылуу изилдөөдө, сөздүн жана тексттин бул өңүттөгү типологиялык көрсөткүчтөрүн (индекстерин) иштеп чыгууда жана сөз формалардын орточо кайталанышын, жыштык сөздүктүн ар кандай участкаларынын текстти камтуу өзгөчөлүктөрүн териштирип чыгуу зарылдыгы туулат.

Тексттин жалпы көлөмүн анда колдонулган ар түрдүү сөз формаларынын (же сөздөрдүн) санына бөлсөк, мындан алынган чоңдук сөз формасынын (же сөздүн) **орточо кайталанышы** деп аталат да, мазмун жактан ал тилдин морфологиялык системасынын (же сөздүк курамынын) текст уюштуруу мүмкүнчүлүктөрүнүн канчалык өлчөмдө жүзөгө ашырылгандыгын мүнөздөйт.

Бул көрсөткүчтү төмөнкү формула аркылуу аныктоого болот:

$$F_{орт}=\frac{N}{L_{с/ф}}$$

Мында $F_{орт}$- тексте колдонулган сөз формаларынын орточо жыштыгы;

N- тексттин көлөмү, б.а. анда колдонулган сөз формаларынын жалпы жыштыгы;

$L_{с/ф}$- тексте колдонулган ар түрдүү сөз формаларынын саны.

Ошентип, бул көрсөткүч тексте колдонулган сөз жана анын грамматикалык формаларынын байлыгын сыпатоого мүмкүнчүлүк түзөт.

Сөз жана сөз формаларынын орточо жыштыгын текстти мүнөздөөчү квантитативдик типологиялык критерий катары да кароого болот. Аны үчүн төмөнкүдөй эки шарттын бирдей аткарылышы зарыл:

биринчиден, бирдей көлөмдөгү тексттер салыштырылыш керек,

экинчиден, салыштырылып жаткан тексттер, кайсы гана тил үчүн болсо да, бирдей жанрга же стилге жатыш керек.

Квантитативдик типологиялык жактан өзүнчө критерий катары төмөнкү учурлар каралышы шарт:

биринчиден, иликтөөгө алынган тилдер үчүн F чоңдугун салыштыруу, экинчиден, адабий тилдин ар кандай жанры үчүн F чоңдугун салыштыруу, үчүнчүдөн, тексттин көлөмүнүн өсүш темпине карай F-тин өсүш темпин аныктоо.

1,2,3,4,5 таблицаларда ар кандай жанрдагы тексттерде катталган сөз формаларынын орточо кайталанышы боюнча өз ара салыштырылды. Салыштыруу адегенде көлөмү 10 миң сөз колдонуштан турган тексттен башталып (1-табл.) андан соң ирети менен көлөмү 20 миң (1-табл.), 30 миң (2-табл.), 50 миң (2-табл.) жана 100 миң (3-табл.) сөз колдонушту камтыган тексттер боюнча улантылды.

Атап айтканда, мындай көлөмдөгү кыргыз тексттеринин ичинен сөз формаларынын орточо кайталанышы боюнча биринчи позицияда Т.Касымбековдун “Сынгын кылыч” тарыхый романы турса, соңку өңүттү А.Койчиевдин Мисмилдирик романы ээлейт. Иликтөөгө алынган калган тексттер төмөнкүдөй барабарсыздыкка туура келет:

$$F_{VI}<F_{X}<F_{XIII}<F_{V}<F_{XI}<F_{III}<F_{I}<F_{IV}<F_{II}<F_{IX}<F_{VII}<F_{XII}<F_{XIV}<F_{VIII}$$

1-таблица. Сөз формаларынын орточо кайталанышы (N=10-20 миң)

Жыштык сөздүктөр	Стиль же жанр	10000		20000	
		$L_{сф}$	F	$L_{сф}$	F
I.	Журнал	4484	2,230	7680	2,604
II.	Журнал	4248	2,354	7164	2,792
III.	Журнал	4575	2,186	7537	2,654
IV.	Журнал	4258	2,349	7368	2,714
V.	Журнал	4737	2,111	8211	2,436
VI.	Роман	4983	2,007	8278	2,416
VII.	Роман	4136	2,418	6726	2,974
VIII.	Роман	3749	2,667	6111	3,273
IX.	Роман	4196	2,383	6872	2,910
X.	Роман	4954	2,019	8220	2,433
XI.	Роман	4737	2,111	7766	2,575
XII.	Роман	4008	2,495	6335	3,157
XIII.	Роман	4896	2,042	7757	2,578
XIV.	Роман	3967	2,521	5987	3,341

Көлөмү 20 миң сөз колдонуштан турган тексте ар бир сөз формасы орточо эсеп менен 2,7755 жолу кайталанган, б.а. F=2,7755. Ал эми ар бир текст боюнча бул көрсөткүч төмөнкү барабарсыздык боюнча такстыкталат (1-табл.):

$$F_{VI}<F_{X}<F_{V}<F_{XI}<F_{XIII}<F_{I}<F_{III}<F_{IV}<F_{II}<F_{IX}<F_{VII}<F_{XII}<F_{VIII}<F_{XIV}$$

30 миң сөз колдонуштагы тексттердин орточо кайталанышы F=3.08, ар бир текст боюнча ал төмөнкү туюнтмага жооп берет (2-табл.):

$$F_{VI}<F_{X}<F_{V}<F_{III}<F_{XI}<F_{IV}<F_{II}<F_{I}<F_{XIII}<F_{IX}<F_{VII}<F_{XII}<F_{VIII}$$

2-таблица. Сөз формаларынын орточо кайталанышы (N=30- 50 миң)

Жыштык сөздүктөр	Стиль же жанр	30000		50000	
		$L_{сф}$	F	$L_{сф}$	F
I.	Журнал	10032	2,990	13843	3,612
II.	Журнал	10051	2,985	15090	3,313
III.	Журнал	10901	2,752	16568	3,018
IV.	Журнал	10139	2,959	15256	3,277
V.	Журнал	10920	2,747	15926	3,140
VI.	Роман	11254	2,666	-	-
VII.	Роман	8732	3,436	12210	4,095
VIII.	Роман	8027	3,737	11427	4,376
IX.	Роман	8800	3,409	12718	3,931
X.	Роман	11037	2,718	-	-
XI.	Роман	10302	2,912	14329	3,489
XII.	Роман	8198	3,659	11294	4,427
XIII.	Роман	9971	3,009	14133	3,538
XIV.	Роман	-	-	-	-

50 миң сөз колдонушта болгон тексттер үчүн сөз формасынын орточо кайталанышы F=3,66, ал эми ар бир текстке карата анын мааниси төмөнкүдөй өзгөрөт (2-табл.):

$F_{III}<F_{V}<F_{IV}<F_{II}<F_{XI}<F_{XIII}<F_{I}<F_{IX}<F_{VII}<F_{VIII}<F_{XII}$

100 миң сөз колдонушка туура келген тексттер үчүн сөз формасынын орточо кайталанышы F=4,52 болот да, мындагы эки тексттин ар биринде сөз формалардын орточо кайталаныштын мааниси төмөнкүдөй бөлүштүрүлөт (3-табл.):

$F_{III}<F_{V}<F_{IV}<F_{II}<F_{I}<F_{XIII}<F_{XI}<F_{VIII}<F_{XII}$

3-таблица. Сөз формаларынын орточо кайталанышы (N=100 миң)

Жыштык сөздүктөр	Стиль же жанр	N	$L_{сф}$	F
I.	Журнал	100000	22159	4,513
II.	Журнал	100000	24876	4,020
III.	Журнал	100000	27348	3,657
IV.	Журнал	100000	25218	3,965
V.	Журнал	100000	25640	3,900
VI.	Роман	-	-	
VII.	Роман	-	-	-
VIII.	Роман	100000	18141	5,512
IX.	Роман	-	-	-
X.	Роман	-	-	-
XI.	Роман	100000	21752	4,597
XII.	Роман	100000	16759	5,967
XIII.	Роман	100000	22023	4,541
XIV.	Роман	-	-	-

Ошентип, иликтөөгө алынган толук тексттер боюнча сөз формасынын орточо кайталанышы жагынан биринчи орунда Шоокум журналынын тексти турса, соңку орунду Т.Касымбековдун “Сынган кылыч” романы ээлейт (4-табл.). Калган тексттердин көрсөткүчү өйдөкүлөрдүн көрсөткүчтөрүнүн ортосунан орун алат, б.а.:

$F_{VI}<F_{X}<F_{XIV}<F_{IX}<F_{XII}<F_{XI}<F_{XIII}<F_{IV}<F_{VIII}<F_{XII}<F_{II}<F_{I}<F_{III}<F_{V}$

4-таблица. Сөз формаларынын орточо кайталанышы (толук тексттер)

Жыштык сөздүктөр	Стиль же жанр	N	$L_{сф}$	F
I.	Журнал	340846	48410	7,041
II.	Журнал	776535	122132	6,358
III.	Журнал	730756	98557	7,415
IV.	Журнал	577891	93905	6,154

V.	Журнал	1028582	107571	9,562
VI.	Роман	41059	14484	2,835
VII.	Роман	60763	13869	4,381
VIII.	Роман	132015	21440	6,157
IX.	Роман	56592	13917	4,066
X.	Роман	49897	16275	3,066
XI.	Роман	159719	29095	5,490
XII.	Роман	114008	18253	6,246
XIII.	Роман	177672	31085	5,716
XIV.	Роман	24221	6792	3,566

Түрк жана индоевропа тилдери боюнча көлөмү 100 миң сөз колдонгон публицистикалык тексттердин сөз формаларынын кайталанышын салыштырсак, кыргыз, казак, каракалпак, түрк, өзбек тилдеринин гезит тексттери, андан алынган сөздүктөр (Ахабаев, 1969:562-567 ; Айтымбетов, Жетишеков, 1980:156-158; Бабанаров, 1981:17; Мухамедов, 1986:72-139) көрсөткөндөй, сөз формаларынын орточо кайталанышы 4,07 экендигин, ал эми журнал тексгери 3,98 боло тургандыгын тастыктайт.

Ушундай эле шартта кыргыз жана түрк публицистикалык тексттери аталган көрсөткүчтүн 3,74 жана 4,29 экендигин айгинелейт. Румын (молдован) тексттеринде бул көрсөткүч 5,86 жана 5,69 түзөт.

Көлөмү 200 миң сөз колдонушка туура келген каракалпак тексттеринде сөз формаларынын орточо кайталанышы 4,84 болсо, өзбек тексттеринде бул көрсөткүч бир кыйла жогору (F=5,83) турат.

Флективдүү-аналитикалык тилдерде сөз формалары үчүн жогорудагыдай эле көрүнүшкө кириптер болобуз. Мисалы, немис публицистикалык текстгинде ар бир сөз форма орто эсеп менен 7,07 жолу кайталанса, молдован публицистикалык текстгинде ал 7,74 ирет кайталанган.

Ошентип, тектеш түрк (кыргыз, каракалпак, казак, өзбек, түрк) тилдеринин публицистикалык тексттеринде бул көрсөткүчтүн маанисн флективдүү-синтетикалык жана флективдүү-аналитикалык индоевропа тилиндегиге караганда төмөнүрөөк турат.

Сөз формаларынын орточо кайталаныш темпи тексттин көлөмүнүн чоцоюшуна жараша өзгөрүп турары жеке эле тигил же бул тилдин типологиялык бөтөнчөлүктөрүнө көз каранды болбостон, бир эле тилдин түрдүү стиль же жанрына да байланыштуу болот. Бул кубулуш 5 таблица аркылуу даана байкалат.

5-таблица. Кыргыз текстинин көлөмүнө (N) карай сөз формаларынын орточо кайталанышынын (F) өсүшү.

Жыштык сөздүктөр	N(миң сөз колдонуш)					
	10	20	30	50	100	Толук текст
I.	2,230	2,604	2,990	3,612	4,513	7,041
II.	2,354	2,792	2,985	3,313	4,020	6,358
III.	2,186	2,654	2,752	3,018	3,657	7,415
IV.	2,349	2,714	2,959	3,277	3,965	6,154
V.	2,111	2,436	2,747	3,140	3,900	9,562
VI.	2,007	2,416	2,666	-	-	2,835
VII.	2,418	2,974	3,436	4,095	-	4,381
VIII.	2,667	3,273	3,737	4,376	5,512	6,157
IX.	2,383	2,910	3,409	3,931	-	4,066
X.	2,19	2,433	2,718	-	-	3,066
XI.	2,111	2,575	2,912	3,489	4,597	5,490
XII.	2,495	3,157	3,659	4,427	5,967	6,246
XIII.	2,042	2,578	3,009	3,538	4,541	5,716
XIV.	2,521	3,341	-	-	-	3,566

Ошентип, кыргыз тилинде тексттин көлөмүнүн өсүш темпи анда колдонулган сөз формаларынын орточо кайталаныш темпинин ошондой эле өлчөмдө өсүшүнө алып келбейт. Ошол эле учурда флективдүү-аналитикалык англис, румын(молдован) жана башка индоевропа тилдеринде, жогоруда белгилегендей, сөз формасынын орточо кайталанышы тексттин көлөмүнө карай тез темп менен өсөт.

Кыргыз жана башка бардык түрк тилдеринде лексикалык бирдиктердин туруктуу жыштыктарын алыш үчүн тектеш эмес тилдерге караганда чоңураак көлөмдөгү тексттерди иликтөө зарыл экендигин биздин маалыматтар ачык көрсөтөт.

Адабияттар

1. Ахабаев А.А. Алфавитно-частотный словарь языка современных казахских газет // Статистика казахского текста. -Алма-Ата: Наука, 1973. с. 344-464.
2. Ахматов Т.К., Жетекишов М. Структура частотного словаря подъязыка киргизской публицистики [на материале газет за. 1977-1978 г.г.) // Материалы семинара "Статистическая оптимизация преподавания языков и инженерная лингвистика", -Чимкент: Чимкентский педагогический институт, 1980, с.156-158.
3. Бабанаров А. Частотный словник и автоматический словарь для машинного перевода турецких газетных текстов // Инженерная лингвистика и оптимазиция преподавания иностранных языков. Л.,1980. с.48-55.
4. Жубанов А.К. Основные принципы формализации содержания казахского текста. Алматы, 2002, 250 с.
5. Мухамедов С.А. Алфавитно-частотный словарь узбекского языка. –Ташкент: ФАН, 1982. –110с.
6. Мухамедов С.А., Пиотровский Р.Г. Инженерная лингвистика. и опыт системно-статистического исследования узбекских текстов. -Ташкент: ФАН, 1986. -160 с.
7. Садыков Т. Проблемы моделирования тюркской морфологии. – Фрунзе: Изд-во «Илим», 1987. –120 с.
8. Садыков Т., Шаршембаев Б. Манас; Кыргызча-Түркчө чоң көрсөткүч сөздүк. Анкара,2011.-1647 б.