УДК 001.103:002

АЙРЫМ ЖАНРДАГЫ КЫРГЫЗ ТЕКСТ КОРПУСТАРЫНДА СЕЙРЕК КОЛДОНУЛГАН СӨЗ ФОРМАЛАРЫНЫН ТЕКСТТИ КАМТУУ КӨРСӨТКҮЧҮ.

Б.Шаршембаев, А.Ибрагимов
Кыргыз-Түрк "Манас" Университети
i.adilets@gmail.com

Верификация редко встречаемых словоформ в некоторых жанрах корпуса кыргызского текста для различных объемов выборок.

Verification of rare words in certain genres corps Kyrgyz text for different sample sizes.

Инженердик лингвистиканын кырк жылдан ашуун мезгилдеги өнүгүү тажрыйбасы көрсөткөндөй, азыркы учурда машина котормосу, автоматтык индекстөө, аннотация жана реферат түзүү иштерин квантитативдик лингвистиканын методдорун колдонбой жана тигил же бул тилдин жалпы информациялык базасын түзбөй жүзөгө ашыруу мүмкүн эмес.

Статистикалык лексикографиянын бир катар теориялык жана практикалык маселелерин чечүүдө тигил же бул тил үчүн белгилүү бир стиль же жанрдагы текстте эң жыш, орто жыш жана сейрек колдонгон сөз

жана сөз формаларын бөлүп алуу, андан соң алардын текстти камтуу өзгөчөлүктөрүн аныктоо талап кылынат.

Бул маселе төмөнкүчө чечилет:

1. Сейрек жана жыш колдонулган лексикалык бирдиктердин ортосундагы чек аныкталат. Буга чейин топтолгон тажрыйбага таянып, бир жана эки жолу колдонулган сөз формаларын бириктирип сейрек колдонулган сөз формаларынын катарына жаткырабыз да, калган бардык сөз формаларын бириктирип сейрек эмес, жыш колдонулган сөз формалары деп тастыктайбыз.

2. Андан соң сейрек эмес, жыш колдонулган сөз формаларынын текстти камтуу көрсөткүчүн төмөнкүчө аныктайбыз:

$$\xi = \left[1 - \frac{n(1,2)}{N}\right] \cdot 100\%$$

Мында: N – тексттин көлөмү;

n(1,2) –бир жана эки жолу колдонулган сөз формаларынын жалпы саны.

Аталган маселени ишке ашырып жана иликтөө үчүн төмөнкү тексттер 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

Ошентип, тектеш жана тектеш эмес тилдер боюнча бул көрсөткүч өйдөкү формула менен аныкталып, андан алынган натыйжалар төмөнкү таблицаларда сыпатталды.

1-таблица. Кыргыз текстеринде сейрек колдонулган сөз формаларынын квантитативдик мүнөздөмөсү (тексттин көлөмү 10 миң сөз колдонуш)

Жыштык сөздүк	Жанр	N	$L_{еф}$	F_1	f_1%	$F_{1,2}$	$f_{1,2}$%	ξ
I.	Журнал	10000	4484	3098	69,090	3746	83,541	62,54
II.	Журнал	10000	4248	2980	70,151	3623	85,287	63,77
III.	Журнал	10000	4575	3270	71,475	3867	84,525	61,33
IV.	Журнал	10000	4258	2953	69,352	3542	83,185	64,58

V.	Журнал	10000	4737	3339	70,488	4014	84,737	59,86
VI.	Роман	10000	4983	3608	72,406	4292	86,133	57,08
VII.	Роман	10000	4136	2845	68,786	3437	83,100	65,63
VIII.	Роман	10000	3749	2585	68,952	3094	82,529	69,06
IX.	Роман	10000	4196	2916	69,495	3506	83,556	64,94
X.	Роман	10000	4954	3581	72,285	4267	86,132	57,33
XI.	Роман	10000	4737	3293	69,517	3974	83,893	60,26
XII.	Роман	10000	4008	2686	67,016	3289	82,061	67,11
XIII.	Роман	10000	4896	3517	71,834	4154	84,845	58,46
XIV.	Роман	10000	3967	2585	65,163	3172	79,960	68,28

Көрүнүп тургандай, сейрек колдонулган сөз формаларын ылгап алуу үчүн чыгырмалар 10 миң сөз колдонушту (1-табл.) түзгөн тексттен баштапып, 20 миң (2-табл.), 30 миң (3-табл.), 50 миң (4-табл.), 100 миң (5-табл.) жана бардык сөз колдонушту (6-табл.) камтыган тексттерге чейин улам көлөмү чоңойтулуп олтурду. Мында көлөм чоңойгон сайын анда колдонулган сөз формаларынын саны да 3749дан тартып (1-табл.) 122132га чейин (6-табл.) бир калыпта өскөндүгү катталды. Ошол эле учурда бир ирет колдонулган сөз формаларынын саны F_1 мамычасында, анын ошол тексттеги сөз формаларынын жалпы санына карата үлүшү f_1 мамычасында, бир жана эки ирет колдонулган сөз формаларынын жалпы санынын кошундусу $F_{1,2}$ мамычасында, алардын өйдөкүдөй үлүшү $f_{1,2}$ мамычасында, акырында сейрек эмес, жыш колдонулду деп эсептелген сөз формаларынын үлүшүн ξ мамычасында берилди.

2-таблица. Кыргыз текстеринде сейрек колдонулган сөз формаларынын квантитативдик мүнөздөмөсү (тексттин көлөмү 20 миң сөз колдонуш)

Жыштык сөздүк	Жанр	N	$L_{сф}$	F_1	f_1%	$F_{1,2}$	$f_{1,2}$%	ξ
I.	Журнал	20000	7680	5178	67,422	6288	81,875	68,56
II.	Журнал	20000	7164	4763	66,485	5777	80,639	71,115
III.	Журнал	20000	7537	5115	67,865	6137	81,425	69,315
IV.	Журнал	20000	7368	4885	66,300	5961	80,904	70,195
V.	Журнал	20000	8211	5542	67,495	6711	81,732	66,445
VI.	Роман	20000	8278	5624	67,939	6886	83,184	65,57
VII.	Роман	20000	6726	4388	65,239	5381	80,003	73,095
VIII.	Роман	20000	6111	4089	66,912	4940	80,838	75,3
IX.	Роман	20000	6872	4489	65,323	5504	80,093	72,48
X.	Роман	20000	8220	5599	68,114	6806	82,798	65,97
XI.	Роман	20000	7766	5038	64,873	6199	79,822	69,005
XII.	Роман	20000	6335	4026	63,552	4959	78,279	75,205
XIII.	Роман	20000	7757	5140	66,263	6294	81,140	68,53
XIV.	Роман	20000	5987	3581	59,813	4518	75,464	77,41

Ошентип, салыштырылып жаткан чыгармалардын бардык тексттеринин көлөмү орто эсеп менен эки эселенгендигине карабай, F_1, f_1, $F_{1,2}$, $f_{1,2}$ мамычаларында көрсөтүлгөндөй, сейрек кездешкен сөз формаларынын үлүш жагынан өсүш темпи барган сайын көбөйбөстөн, тескерисинче, азаюунун үстүндө болот. Сал.: Жаңы Ала-Тоо журналы, 2013 ж:

f_1=69,352<66,300<65,638<65,083<61,888<58,170%;

$f_{1,2}$=83,185<80,904<79,998<79,339<76,612<72,074%.

Демек, мында тексттин көлөмү чоңойгон сайын андагы сейрек колдонулган сөз формаларынын саны да азайып олтурары байкалат. Бул-кыргыз жана башка тектеш түрк тилдерине мүнөздүү бөтөнчөлүк (сал.:1-6 табл., f_1,$f_{1,2}$ мамычаларын).

3-таблица. Кыргыз текстеринде сейрек колдонулган сөз формаларынын квантитативдик мүнөздөмөсү (тексттин көлөмү 30 миң сөз колдонуш)

Жыштык сөздүк	Жанр	N	$L_{сф}$	F_1	f_1%	$F_{1,2}$	$f_{1,2}$%	ξ
I.	Журнал	30000	10032	6490	64,693	8006	79,805	73,313
II.	Журнал	30000	10051	6555	65,217	8066	80,251	73,113
III.	Журнал	30000	10901	7357	67,489	8861	81,286	70,463

IV.	Журнал	30000	10139	6655	65,638	8111	79,998	72,963
V.	Журнал	30000	10920	7109	65,101	8745	80,082	70,850
VI.	Роман	30000	11254	7437	66,083	9160	81,393	69,467
VII.	Роман	30000	8732	5492	62,895	6794	77,806	77,353
VIII.	Роман	30000	8027	5208	64,881	6359	79,220	78,803
IX.	Роман	30000	8800	5550	63,068	6828	77,591	77,240
X.	Роман	30000	11037	7312	66,250	8936	80,964	70,213
XI.	Роман	30000	10302	6474	62,842	8032	77,965	73,227
XII.	Роман	30000	8198	5026	61,308	6293	76,763	79,023
XIII.	Роман	30000	9971	6332	63,504	7837	78,598	73,877
XIV.	Роман	30000	-	-	-	-	-	-

4-таблица. Кыргыз тексгеринде сейрек колдонулган сөз формаларынын квантитативдик мүнөздөмөсү (тексттин көлөмү 50 миң сөз колдоиуш)

Жыштык сөздүк	Жанр	N	$L_{сф}$	F_1	f_1%	$F_{1,2}$	$f_{1,2}$%	ξ
I.	Журнал	50000	13843	8550	61,764	10697	77,274	78,606
II.	Журнал	50000	15090	9641	63,890	11943	79,145	76,114
III.	Журнал	50000	16568	10864	65,572	13225	79,823	73,550
IV.	Журнал	50000	15256	9929	65,083	12104	79,339	75,792
V.	Журнал	50000	15926	10021	62,922	12443	78,130	75,114
VI.	Роман	50000	-	-	-	-	-	-
VII.	Роман	50000	12210	7341	60,123	9164	75,053	81,672
VIII.	Роман	50000	11427	7198	62,991	8837	77,334	82,326
IX.	Роман	50000	12718	7871	61,889	9736	76,553	80,528
X.	Роман	50000	-	-	-	-	-	-
XI.	Роман	50000	14329	8696	60,688	10881	75,937	78,238
XII.	Роман	50000	11294	6696	59,288	8399	74,367	83,202
XIII.	Роман	50000	14133	8642	61,148	10875	76,948	78,250
XIV.	Роман	50000	-	-	-	-	-	-

5-таблица. Кыргыз тексгеринде сейрек колдонулган сөз формаларынын квантитативдик мүнөздөмөсү (тексттин көлөмү 100 миң сөз колдонуш)

Жыштык сөздүк	Жанр	N	$L_{сф}$	F_1	f_1%	$F_{1,2}$	$f_{1,2}$%	ξ
I.	Журнал	100000	22159	13211	59,619	16531	74,602	83,469
II.	Журнал	100000	24876	15203	61,115	19056	76,604	80,944
III.	Журнал	100000	27348	17156	62,732	21279	77,808	78,721
IV.	Журнал	100000	25218	15607	61,888	19320	76,612	80,680
V.	Журнал	100000	25640	15475	60,355	19381	75,589	80,619
VI.	Роман	100000	-	-	-	-	-	-
VII.	Роман	100000	-	-	-	-	-	-
VIII.	Роман	100000	18141	10980	60,526	13590	74,913	86,410
IX.	Роман	100000	-	-	-	-	-	-
X.	Роман	100000	-	-	-	-	-	-
XI.	Роман	100000	21752	12486	57,402	15681	72,090	84,319
XII.	Роман	100000	16759	9440	56,328	11932	71,198	88,068
XIII.	Роман	100000	22023	12854	58,366	16253	73,800	83,747
XIV.	Роман	100000	-	-	-	-	-	-

Эми аталган тексттерде колдонулган сөз формаларынын ичинен сейрек колдонулган сөз формаларынын үлүшү кандай болду экен деген суроо койсок, анда анын жообун төмөнкүчө берүүгө болот: жыштык сөздүктө катталган өйдөкү сөз формаларынын жүздөн алтымыш беши, негизинен, сейрек колдонулган сөз формаларынын үлүшүнө туура келет. Кыскасы, каралган тексттерде колдонулган ар түрдүү сөз формаларынын теңинен көбүн дал ушул сейрек колдонулган сөз формалары түзөт. Бул факт кыргыз тилиндеги тексттер сөз формасы бай экендигин кадиксиз тастыктайт.

6-таблица. Кыргыз текстеринде жыш колдонулган сөз формаларынын квантитативдик мүнөздөмөсү (толук тексттер үчүн)

Жыштык сөздүк	Жанр	N	$L_{сф}$	F_1	f_1%	$F_{1,2}$	$f_{1,2}$%	ξ
I.	Журнал	340846	48410	26547	54,838	33889	70,004	90,057
II.	Журнал	776535	122132	68637	56,199	88590	72,536	88,592
III.	Журнал	730756	98557	49278	49,999	67741	68,733	90,730
IV.	Журнал	577891	93905	54625	58,170	67681	72,074	88,288
V.	Журнал	1028582	107571	56458	52,484	72422	67,325	92,959
VI.	Роман	41059	14484	9298	64,195	11523	79,557	71,936
VII.	Роман	60763	13869	8203	59,146	10293	74,216	83,060
VIII.	Роман	132015	21440	12571	58,633	15815	73,764	88,020
IX.	Роман	56592	13917	8543	61,385	10594	76,123	81,280
X.	Роман	49897	16275	10512	64,590	12915	79,355	74,117
XI.	Роман	159719	29025	16238	55,945	20484	70,574	87,175
XII.	Роман	114008	18253	10202	55,892	12946	70,925	88,645
XIII.	Роман	177672	31085	17422	56,046	22125	71,176	87,547
XIV.	Роман	24221	6792	4028	59,305	5078	74,764	79,035

Эми чыгырмаларыбыздын түрдүү тексттеринде жыш колдонулган сөз формаларынын колдонуш үлүшүнө (сал. ξ мамычасындагы белгилерди) көңүл бура турган болсок, тексттин көлөмүнө карай бул көрсөткүчтүн басаңдаган темп менен бир калыпта өскөндүгүн байкайбыз (1-6 табл.). Маселен, мында Сынгын кылыч тарыхый романы төмөнкүлөрдү көрсөтөт:

ξ=62,54<68,56<73,313<78,606<83,469<90, 057%.

Тактап айтканда, бул варианттын алгачкы көлөмүнө (N=10 миң сөз колдонуш) караганда кийинки көлөмүндө (N=20 миц сөз колдонуш) жыш колдонулган сөз формаларыны үлүшү ≈6%га өссө, андан кийинкиде (N=30 миң сөз колдонуш) мурункусуна (N=20 миң сөз колдонуш) караганда ≈5%га, кийинкиде (N=50 миң сөз колдонуш) ≈5%га, акыркысында (N=100 миң сөз колдонуш) ≈6%га өскөн. Калган тексттердин көрсөткүчү да ушундай эле мүнөзгө ээ. Демек, бул фактылар тексттин көлөмү өскөн сайын анда колдонулган жогорку жыштыктагы сөз формаларынын текстти камтуу мүмкүнчүлүгү турукташа тургандыгын жана мындай турукташуу тилдин табиятына шайкеш келерин билдирет. Ошондуктан мындай учурдан алынган натыйжалар статистикалык жактан да, лингвистикалык жактан да илимий иликтөөлөргө коюлуучу чен өлчөмдөргө (критерийлерге) ар дайым толук жооп берет. Квантитативдик иликтөөнүн башкы максаты да дал ушунда.

Чыгырмалар тексттеринин ичинен өйдөкү көрсөткүч боюнча салыштырууга көлөмдөрү белгилүү өлчөмдө бирдей болгон романдар төмөнкүдөй тартипте жайгашат (6-табл.):

Мисмилдирик(Койчиев)>Башы жок чабендес(М.Рид)>Өмүр кызык(Лондон)

(ξ=88,02 ξ=87,547 ξ=87,175)

Бул үч текстте сейрек колдонулган сөз формаларынын саны чоңойот.

Албетте, кыргыз тилиндеги текст корпустарын жакын арада кол менен териштирип чыгуу мүмкүн эмес. Ошондуктан масштабдуу лексикографиялык, стилистикалык, грамматикалык, социолингвистикалык маселелерди чечиш үчүн түркологияда соңку муундагы кубаттуу ЭЭМдерди колдонуу маселеси күн тартибинде турат.

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.