УДК 811.512.145'33

Җ.Ш. Сөләйманов,
Б.Э. Хәкимов,
Р.А. Гыйльмуллин

ТАТАР ТЕЛЕ КОРПУСЫН ТӨЗҮНЕҢ КОНЦЕПТУАЛЬ ҺӘМ ЛИНГВИСТИК АСПЕКТЛАРЫ

Статья посвящена вопросам разработки концептуальной основы и уточнения модели корпуса татарского языка. Исследуются особенности представления лингвистической информации и предлагаются принципы корпусной морфологической разметки текстов на татарском языке. Самостоятельную проблему представляет обеспечение репрезентативности текстовой коллекции корпуса, одним из путей решения которой является статистический анализ лингвистических особенностей текстов. На процесс разработки корпуса татарского языка и его лингвистической разметки влияют особенности языковой системы, в частности соотношение регулярных и нерегулярных явлений.

Ключевые слова: татарский язык, корпусная лингвистика, информационные технологии, лингвистические модели, репрезентативность корпуса текстов, лингвистическая корпусная разметка.

Табигый телләрне эмпирик төгәл материал нигезендә тел мәгълүматларын автомат рәвештә эшкәртү технологияләрен кулланып тикшерү хәзерге фән өлкәсендә перспектив юнәлешне тәшкил итә. Телләрнең электрон корпусларын төзү күп кенә лингвистик мәсьәләләрне чишүнең иң нәтиҗәле чарасы санала. Татар теле өчен шундый системаны булдыру тел төзелеше, аның лексик составы турында яңа мәгълүматлар алу белән бергә башка телләрнең, аерым алганда төрки телләрнең, лингвистик модельләрен төзү, әлеге телләрдәге текстларны электрон технологияләр ярдәмендә эшкәртү өлкәсендә дә яңа мөмкинлекләр тудыра.

Татар теле корпусын төзү өлкәсендә тикшеренүләр XX гасырның 90 нчы елларыннан башлап Татарстан Республикасы Фәннәр академиясе һәм Казан (Идел буе) федераль университетындагы фәнни коллективлар тарафыннан алып барыла [Бухараев, Сулейманов, 1995; Бухараев, Сафиуллина, Сулейманов, 1995]. Аның төп юнәлешләрен компьютерлар өчен лингвистик модельләрнең формаль алгоритмнарын эшләү һәм машина тәрҗемәсе, эзләү системалары, компьютер лексикографиясе кебек өлкәләрне колачлаган гамәли тикшеренүләр тәшкил итә. Алга таба машина фонды концепциясе билгеле бер үзгәрешләр кичереп һәм тулыланып, татар теленең барлык лингвистик ресурсларын колачлаган гомумтел корпусы концепциясенең нигезен тәшкил итте. Хәзерге вакытта шундый гомумтел корпусы концепциясе «Туган тел» корпусы проектында гамәлгә ашырыла [Татарский национальный корпус...].

1. Татар теле корпусының концептуаль моделе. Мәгълүм булганча, һәр табигый тел кешелек җәмгыятендә аралашу һәм дөньяны тасвирлау максатларына хезмәт итә торган катлаулы семиотик системаны тәшкил итә. Бүгенге көндә дөньяның үзенчәлекле информацион моделе саналган телнең теге яки бу күренешләре модельләрен төзүнең төрле юллары бар. Телне формаль яктан тулаем тасвирлау мөмкин булмау сәбәпле, безнең тарафтан лингвистик модельләрне төзүдә билгеле бер мәсьәләләрне хәл итү өчен «контекстуаль тулылык» принцибы нигезендә минималь күләмдәге чараларны билгели торган прагматик юнәлеш тәкъдим ителде. Нәтиҗәдә прагматик юнәлештәге лингвистик модельләрнең түбәндәге төрләре аерып чыгарылды:

1) контекстның эченә үтеп керергә, аны трансформацияләргә мөмкинлек биргән, лингвистик модельләрдән башка «контекст моделе» һәм «дөнья сурәте моделе» кебек экстралингвистик модельләрне дә кулланган семиотик модельләр;

2) автоматлаштырылган система белән кеше арасында табигый телдә аралашуны тәэмин иткән, үз эченә коммуникатив һәм контекстуаль модельләрне, минималь лингвистик информацияне алган интерактив модельләр;

3) текстларны телнең билгеле бер тармакларына караган формаль кагыйдәләр нигезендә максатчан эшкәртүне тәэмин иткән концептуаль-формаль модельләр;

4) табигый тел тармакларының мөһим берәмлекләренең универсаль тасвирламасын тәшкил иткән концептуаль-функциональ модельләр [Сулейманов, Гатиатуллин, 2003].

Әлеге классификация нигезендә татар теле корпусы телнең төрле тармакларының үзара тыгыз бәйләнгән концептуаль-функциональ модельләре комплексы буларак күзаллана. Ачык система буларак ул башка лингвистик модельләрне һәм алар белән бәйле автоматлаштырылган тамгалау системаларын да өстәргә мөмкинлек бирә. Ягъни корпус үз эчендә морфологик, синтаксик, семантик һ.б. модельләрне берләштерә.

Электрон корпусның информацион моделе үз эченә морфологик, синтаксик, семантик һ.б. лингвистик мәгълүматны ала. Әлеге мәгълүмат типлары үзара тыгыз бәйләнештә тора. Алар мәгълүматны автомат рәвештә эшкәртүне тәэмин итүче махсус формаль күрсәткечләр (тамгалау системалары) белән билгеләнә торган лингвистик параметрлар рәвешендә тәртипкә салынган. Лингвистик мәгълүмат параметрлары корпус берәмлекләрендә чагылыш таба.

Шуның белән бергә корпус моделе мәгълүматны эшкәртү системасыннан, үз эченә тамгаланган текстлар коллекциясен алган мәгълүмат базаларыннан һәм алар белән идарә итү системасыннан оешкан технологик схеманы тәшкил итә.

2. Мәсьәләне хәл итү юлы буларак корпусның репрезентативлыгы. Репрезентативлык, ягъни җитди лингвистик нигез, барлык корпусларның да мөһим

үзенчәлеге һәм аларны эшләүнең зарури шарты булып тора. Моңа ирешүнең дискурска, телнең яшәешенә яисә тел күренешләре төрлелеген, шул исәптән сирәк һәм пассив тел берәмлекләрен чагылдыруга юнәлтелгән төрле стратегияләре бар [Баранов, 2001; Шимкова, 2005].

Татар теле корпусына килгәндә, текстлар коллекциясе башлангыч этапта ук билгеле бер репрезентативлыкка ия булырга һәм нәтиҗәле тикшеренүләр үткәрү мөмкинлеге тудырырга тиеш. Бу мәсьәләне хәл итү текстларны аларда төрле тел күренешләренең, шул исәптән еш кулланылмый торган берәмлекләрнең дә урын алуын күз алдында тотып сайлаганда гына мөмкин. Моның өчен текстларны корпуска кертер алдыннан аларга җентекле статистик анализ үткәрү сорала.

Әлеге максатта безнең тарафтан гомуми күләме 600000 нән күбрәк сүз кулланылышы тәшкил иткән татарча текстларның эксперименталь коллекциясенең төрле квантитатив һәм лексик-грамматик параметрларының статистик үзенчәлекләре тикшерелде [Гильмуллин, Невзорова, Хакимов, 2011].

Тикшеренү нәтиҗәләреннән күренгәнчә, сүз төркемнәре барлык текстларда да нигездә бер үк кулланылыш ешлыгына ия. Җыелма текстта да аерымлык күзәтелми. Әлеге текстлар төрле күләмгә ия, алар XX гасыр дәвамында төрле вакытта язылганнар. Шул рәвешле түбәндәге таблицада чагылыш тапкан статистик параметрлар сәнгатьле проза стиле өчен хас булган «уртача норманы» билгели торган константалар ролен үтиләр. Нәтиҗәдә конкрет текстларның статистик үзенчәлекләрендә бу константалардан читкә тайпылыш дәрәҗәсен билгеләргә мөмкинлек туа.

Мәсәлән, беренче текстта (Т. Галиуллин хикәяләре), бер яктан, исемнәрнең чагыштырмача югары, сыйфатларның түбән ешлыгы, ә икенче яктан, рәвеш, алмашлык һәм ярдәмлек сүз

1 нче таблица

Сүз төркемнәре статистикасы: сәнгатьле проза (%)

	Исемнәр	Фигыльләр	Сыйфатлар	Рәвешләр	Алмашлыклар	Ярдәмлек сүз төркемнәре
Т. Галиуллин	39,32	35,68	9, 85	1,52	7,4	6,22
Г. Гыйльманов	35,13	36,71	8,02	2,22	9,94	7,99
Ә. Еники	35,05	33,75	9,14	2,59	9,56	9,91
Р. Зәйдулла	35,6	37,08	7,86	2,33	9,27	7,86
Г. Кутуй	35,51	37,03	8,2	2,58	8,03	8,65
Г. Бәширов	34,8	35,76	8,83	2,93	9,58	8,1
Ф. Әмирхан	35,17	35,27	8,52	2,86	9,38	8,8
Җыелма текст	36,09	36,06	8,57	2,29	9	7,99

төркемнәренең аз очравы (2 нче таблица) әлеге текстта аерым тел конструкцияләренең үзенчәлекле кулланылышы турында сөйли. Ул, корпуска кертелгән очракта, төрлелекне арттыра һәм нигездә корпусның репрезентативлыгына да тәэсир итәргә тиеш.

Уку-укыту өлкәсенә караган текстларда шулай ук төрле мәгънәләргә ия булган аерым тенденцияләр күзәтелә. Биредәге мәгънә төрлелеге, безнең фикеребезчә, беренчедән, коллекциядә бу жанрдагы текстларның чагыштырмача аз санда булуы, икенчедән,

2 нче таблица

	Исемнәр	Фигыльләр	Сыйфатлар	Рәвешләр	Алмашлыклар	Ярдәмлек сүз төркемнәре
Беренче текст	36,67	32,17	12,43	1,37	6,2	8,16
Икенче текст	38,73	34,97	10,99	1,34	5,66	8,31
Өченче текст	41,75	39,4	9,15	0,88	3,17	5,65
Җыелма текст	39,74	34,81	11,17	1,26	5,34	7,69

3 нче таблица

Сүз төркемнәре статистикасы: газета текстлары (%)

	Исемнәр	Фигыльләр	Сыйфатлар	Рәвешләр	Алмашлыклар	Ярдәмлек сүз төркемнәре
Беренче текст	40,48	32,21	10,01	1,68	6,75	8,89
Икенче текст	45,3	30,26	10,36	1,34	5,44	7,3
Җыелма текст	42,47	31,4	10,16	1,54	6,2	8,23

4 нче таблица

Сүз төркемнәре статистикасы: жанрлар буенча җыелма мәгълүматлар (%)

	Исемнәр	Фигыльләр	Сыйфатлар	Рәвешләр	Алмашлыклар	Ярдәмлек сүз төркемнәре
Сәнгатьле проза текстлары	36,09	36,06	8,57	2,29	9	7,99
Уку-укыту өлкәсенә караган текстлар	39,74	34,81	11,17	1,26	5,34	7,69
Газета текстлары	42,47	31,4	10,16	1,54	6,2	8,23
Җыелма текст	38,51	35	10,02	1,71	6,91	7,86

аларның төрле «фәннилек» дәрәҗәсенә ия булуы һәм төрле предмет өлкәләренә каравы белән аңлатыла. Шуңа да карамастан, сәнгатьле проза белән чагыштыруганда ачык күренеп торган номинативлык, сыйфатларның активлыгы һәм, киресенчә, алмашлыкларның пассивлыгы жанрның квантитатив маркерлары буларак хезмәт итәргә мөмкин.

3 нче таблицадан күренгәнчә, газета текстлары да аларга хас булган нормадан читкә тайпылышларны һәм тел төрлелеген билгеләргә мөмкинлек бирә торган аерым константаларга һәм жанр-стиль маркерларына ия.

Шул рәвешле, төрле функциональ кулланылышка ия һәм күләме чикләнгән текстлар корпусының репрезентативлыгына билгеле бер дәрәҗәдә аларның квантитатив һәм статистик үзенчәлекләрен анализлау нигезендә ирешергә мөмкин. Әлеге метод репрезентативлык параметрларының тагын да төгәлрәк классификациясен эшләү хисабына камилләштерелергә мөмкин.

3. Татар теле корпусында морфологик тамгалау системасы. Татар телендә сүзнең морфологик төзелешен анык агглютинатив табигате билгели: сүзформалар сүз ясагыч һәм сүз төрләндергеч аффиксларның нигезгә билгеле бер тәртиптә ялгануы нәтиҗәсендә ясалалар. Һәр грамматик мәгънә, кагыйдә буларак, аерым аффикс ярдәмендә белдерелә. Аффикслар даими, ягъни аерым сүз төркемнәренә караган барлык сүзләргә дә кушылалар. Шуңа бәйле рәвештә морфологияне формаль яктан күзаллау өчен аерым очракларда нигезләр сүзлегенә мөрәҗәгать итеп, аффикслар чылбырын анализлау сорала.

3.1. Татар текстларын морфологик тамгалау моделе. Татар теленең бүгенге көндә эшләнгән морфологик модельләре аның агглютинатив табигатен тасвирлаганда билгеле бер стандартларга нигезләнәләр [Suleymanov, Guilmoulline, Guilmoulline, 2000; Сулейманов, Гатиатуллин, 2003]. Татар теле корпусындагы текстларны морфологик тамгалауны автоматлаштыру лингвопроцессорлар тарафыннан кулланылган махсус программа чараларын булдыру юлы белән тормышка ашырыла. Аерым алганда, автоматлаштырылган морфологик тамгалау системасының үзәк компоненты буларак татар теленең ике катлы морфологик моделе нигезендә махсус җайлаштырып эшләнгән анализатор кулланыла [Гильмуллин, 2009; Хакимов, Гильмуллин, 2009].

Татар теле морфологиясенең әлеге моделе сүзформаны билгеле бер тәртиптә урнашкан ике катта чагылышлы аффикслар чылбыры буларак күзаллауга нигезләнә. Беренче («тышкы») катта аффикслар фонологик үзенчәлекләре белән, алломорфлар буларак тасвирланса, икенче, ягъни «эчке», катта шул алломорфларның гомумиләштерелгән берәмлеге, морфема буларак күрсәтелә. Мисал өчен, «эчке» катта морфотактика структурасында аерым билгеле урын алып торган *«-га»* аффиксы-морфемасы, «тышкы» катта *«-га, -гә, -ка, -кә, -а, -ә, -на, -нә»* аффикслары-алломорфлары

белән чагыла. Анализаторның эшчәнлеге фигыль һәм номинатив парадигмалар нигезендә корылган һәм сүз нигезе белән телдәге аффикслар арасындагы үзара бәйләнешләрне билгели торган морфотактик кагыйдәләр базасы нигезендә тормышка ашырыла [Гильмуллин, 2009].

Алга таба татарча текстта морфологик тамгалауны автоматлаштыру мөмкинлекләрен карап үтик. Татар телендәге сүз төркемнәренең грамматик үзенчәлекләре тел белеменә караган күпсанлы хезмәтләрдә тулы һәм җентекле рәвештә тасвирланган [Татарская грамматика, Т. II, 1993]. Аларны тану һәм автомат рәвештә тамгалау түбәндәге формаль күрсәткечләр нигезендә башкарылырга мөмкин:

I. Лексик-грамматик (семантик) төркемчәләрнең ачык күренеп торган формаль күрсәткечләре юк, ә кайберләренә хас булган формаль күрсәткечләр тотрыклы түгел. Шуңа өстәп, ирекле сүзформаның мондый структурасы турыдан-туры билгеле бер семантик төркем белән бәйләнмәгән.

II. Морфологик категорияләр исә, гадәттә, билгеле бер сүзформасын чагылдыра торган ачык формаль күрсәткечләргә ия.

Димәк, татар теленең морфологик үзенчәлекләре сүзформаның аффикслар составы автомат рәвештә анализланган вакытта чагыштырмача җиңел танылалар. Төрле сүз төркемнәренең лексик-грамматик (семантик) төркемчәләрен тану һәм автомат рәвештә тамгалау исә морфологик күрсәткечләр нигезендә генә хәл ителә алмый. Шул сәбәпле корпусны төзүнең башлангыч этабында морфологик тамгалау системасында эзлекле рәвештә билгеле бер аффикслар ярдәмендә белдерелгән грамматик категорияләр турындагы мәгълүмат кына урнаштырылырга мөмкин.

3.2. Татар теле морфологиясендә тотрыксыз күренешләр. Алга таба татар теленең тотрыклы табигатькә ия морфологиясендә урын алган гомум кагыйдәләрдән читкә тайпылыш очракларына һәм аларны текстларны автомат рәвештә тамгалау процессында эшкәртү мөмкинлекләренә тукталу сорала. Тел системасына мөнәсәбәттә бу күренешләрне ике төргә аерырга мөмкин: тышкы (системага карамаган) һәм эчке (системага бәйле) тайпылышлар.

1. Тышкы тайпылышларга ассимиляцияләнмәгән алынмалардагы морфологик үзгәрешләр һәм хәзерге татар орфографиясендәге җитешсезлекләр белән бәйле очраклар керә. Әлеге тайпылышларның төрләре Җ.Ш. Сөләймановның «Регулярность морфологии татарского языка и типы нарушений в языке» мәкаләсендә тасвирланган [Сулейманов, 1994]. Бүгенге көндә хәзерге татар орфографиясе һәм алынмаларны үзләштерү принциплары нигезендә бу төр сүзләрне автомат рәвештә эшкәртү мәсьәләсен хәл итү һәм алардагы үзгәрешләр белән бәйле закончалыкларны тасвирлау чараларын табу сорала.

2. Тышкы тайпылышлардан башка татар теленең морфологик системасында автомат рәвештә

танылырга мөмкин булган эчке тайпылышлар да урын ала. Әлеге тайпылышлар тел берәмлекләре өчен хас универсальлек, полифункциональлек, экономиялелек белән бәйле рәвештә барлыкка килеп, аларның системалы билгеләре саналалар һәм аерым эчке закончалыкларга буйсыналар.

Автомат рәвештә тамгалау вакытында аерым кыенлыкларны полифункциональ һәм омонимик аффикслар, «нуль» формалар, тиңдәш кисәкләр (аффикс аларның соңгысына гына кушылырга мөмкин), билгеле бер категория сүзләрнең, аерым алганда, кайбер алмашлык һәм бәйлек сүзләрнең, төрләнешендәге үзенчәлекләр тудырырга мөмкин.

3.3. Татар теле корпусында морфологик мәгълүматны стандартлаштыру. Табигый телне автомат рәвештә эшкәртү системасы өчен концептуаль һәм технологик стандартларны булдыру зур әһәмияткә ия. Татар теле корпусының морфологик стандарты мәгълүматны бертөрле итеп тәкъдим итәргә, морфологик аннотациянең теоретик нигезен оештырырга һәм үз эченә татар теленә хас булган теге яки бу морфологик күренешләрне, шул исәптән бәхәсле очракларны да, тасвирлау кагыйдәләрен алырга тиеш.

Морфологик мәгълүматны тәкъдим итү структурасына килгәндә, биредә лексемаларның кайсы сүз төркеменә караганлыгын билгеләү һәм сүзформаларның морфологик үзенчәлекләрен аңлату белән бәйле бәхәсле мәсьәләләрне аерып чыгарырга мөмкин.

Татар грамматикасында сүз төркемнәрен тасвирлау өлкәсенә ике төп бәхәсле мәсьәлә хас. Беренчесе аффикслар алмаган хәлдә атрибутив функциядә килгән сүзләрнең кайсы сүз төркеменә караганлыгын билгеләү белән бәйле. Биредә аларны артык омонимия очракларын булдырмас өчен исем сүз төркеменә кертеп карау дөресрәк булыр иде.

Икенчесе күпкә катлаулырак һәм ул бер-берсен алыштырып килгән сүз төркемнәрен (сыйфатлар һәм рәвешләрне, исемнәр һәм сыйфатларны) аеру өлкәсенә карый. Мәсьәләне ике төрле юл белән чишәргә мөмкин: әлеге очракларны сүз төркемнәре омонимиясе итеп карап, сүзләрне контекст ярдәмендә аерырга яисә һәр лексеманы аерым тасвирларга. Ике юлның да, үз өстенлекләре булган кебек, кимчелекләре дә бар һәм алар аерым карауны сорый.

Татар теленең килешләр системасы белән бәйле мәсьәлә шулай ук катлаулы очраклардан санала. Соңгы тикшеренүләр аның хәзерге грамматик теория кысаларында тулаем хәл ителә алмавын күрсәтте [Сулейманов, 1996]. Сүзформаларның аффикслар составы белән бәйле рәвештә әлеге мәсьәләнең ике аспекты аерылып чыга: нуль форманың интерпретациясе һәм кайбер аффиксларның килеш категориясенә карата статусын билгеләү. Безнең фикеребезчә, биредә телнең табигате белән грамматик теория өлкәсендәге традицияләр арасындагы балансны саклау зарур. Мәсәлән, аффикслар алмаган формалар

барлык очракларда да «баш килеш» буларак күрсәтелергә, ә статуслары ахырга кадәр ачыкланып бетмәгән аффикслар шартлы рәвештә килеш парадигмасына кертелергә тиеш.

Тагын бер катлаулы очракны татар фигыле өчен хас юнәлеш категориясе, аерым алганда сүз төрләндергеч һәм сүз ясагыч функцияләрне башкара алган юнәлеш кушымчаларына хас үзенчәлекләр тәшкил итә. Нәтиҗәдә юнәлеш аффиксларын алган фигыльләрне леммалаштыру (нигезен билгеләү) мәсьәләсе туа. Бу очракта альтернатива булдыру сорала һәм нигез сыйфатында бердән артык фигыль формасы күрсәтелә ала. Әлеге алымны инфинитив формаларына карата да кулланырга мөмкин, чөнки биредә дә күп кенә очракларда төрлелек күзәтелә.

3.4. Татар теле корпусында морфологик тамгалауның форматы. Корпуста сүзформасы турындагы морфологик мәгълүмат гомуми планда ике төп «кырдан» оеша: 1) сүз төркеме буларак тасвирламасы; 2) морфологик билгеләр (параметрлар) тупланмасы.

Татар теле өчен хас морфотактика үзенчәлекләрен, бер яктан, катлаулы һәм гади, икенче яктан, мәҗбүри һәм факультатив параметрларга бүлеп карарга мөмкин. Катлаулы параметрлар берничә мәгънәгә ия булып, нинди дә булса грамматик категориягә карыйлар (мәсәлән, килеш кушымчалары системасы). Гади параметрлар исә бер генә мәгънәгә ия булырга мөмкиннәр (мәсәлән, *-мы* сорау формасы). Параметрның мәҗбүрилеге һәм факультативлыгына килгәндә, мәҗбүри параметр билгеле бер сүз төркеменең нинди дә булса сүзформасына карарга тиеш (мәсәлән, исемнәр һәрвакыт нинди дә булса килештә торалар, «килешсез» исем формалар юк). Факультатив параметр, аерым бер конкрет мәгънәдән башка, кире мәгънәне дә алырга мөмкин (мәсәлән, билгенең юклыгы – татар телендә исемнәр тартым мәгънәсен белдермәскә дә мөмкиннәр).

Йомгак. Татар теле корпусының концептуаль моделе һәм татар текстларын морфологик тамгалау системасы алга таба татар телендәге теләсә нинди текстның морфологик үзенчәлекләренең формаль тасвирламасын электрон машина укый ала торган форматта төзергә мөмкинлек бирә. Корпусны эшләүнең башлангыч этабында морфологик тамгалау системасында эзлекле рәвештә билгеле бер аффикслар ярдәмендә белдерелгән грамматик категорияләр турында мәгълүмат бирү белән чикләнү дә җитә. Автомат рәвештә тамгалау системасының нәтиҗәлелеге морфологик анализатор өчен өстәмә фонологик һәм морфотактик кагыйдәләр кертү юлы белән арттырылырга мөмкин.

Хәзерге вакытта түбәндәге мәсьәләләрне хәл итү өлкәсендә эш алып барыла:

1) корпус кысаларында сүзформаларын контекст эчендә анализлау чараларын эшләү;

2) полифункциональ, омонимик аффиксларның, «нуль» формаларның мәгънәләрен билгеләүдә, тиңдәш кисәкләрне та-

нуда автоматик чаралар куллану мөмкинлекләрен ачыклау һәм кагыйдәләрен эшләү;

3) гомум кагыйдәләргә буйсынмаган сүзләр, иң беренче чиратта алынмалар өчен фонологик һәм морфотактик кагыйдәләрне булдыру мөмкинлекләрен тикшерү;

4) төрле аффикс чылбырлары ешлыгын, бер төркемдәге лексемаларга башка парадигма аффиксларының ялгану очракларын корпуслы анализ юлы белән лингвопроцессорларны «өйрәтү» һәм лингвистик модельләргә төзәтмәләр кертү алымнарын һәм чараларын эшләү;

5) татар теле корпусының синтаксик һәм семантик модельләрен, алар белән бәйле тамгалау системаларын эшләү.

Әдәбият

Баранов А.Н. Проблема репрезентативности корпуса данных (на примере политической метафорики) // Труды международного семинара по комьютерной лингвистике и ее приложениям «Диалог-2001». Аксаково, 2001. Т. 2. С. 13–15.

Бухараев Р.Г., Сулейманов Д.Ш. К концепции внедрения татарского языка в компьютерные технологии // Татарский язык и новые информационные технологии. Сер.: Интеллект. Язык. Компьютер. Казань: Казан. гос. ун-т, 1995. Вып. 2. С. 8–19.

Бухараев Р.Г., Сафиуллина Ф.С., Сулейманов Д.Ш. и др. К концепции Машинного Фонда Республики Татарстан // Татарский язык и новые информационные технологии. Сер.: Интеллект. Язык. Компьютер. Казань: Казан. гос. ун-т, 1995. Вып. 2. С.20–35.

Гильмуллин Р.А. Разработка файла морфотактических правил для глагольных групп татарского языка // Проблемы сохранения языка и культуры в условиях глобализации: материалы VII Международного симпозиума «Языковые контакты Поволжья» / науч. ред. И.А. Гилязов. Казань: Казан. гос. ун- т, 2009. С. 222–226.

Гильмуллин Р.А., Невзорова О.А., Хакимов Б.Э. Корпус татарских текстов: проблема репрезентативности // Труды международной конференции «Корпусная лингвистика – 2011». СПб.: С.-Петербург. гос. ун-т, 2011. С. 125–130.

Сулейманов Д.Ш. Регулярность морфологии татарского языка и типы нарушений в языке // Когнитивная и компьютерная лингвистика / науч. ред. Р.Г. Бухараев и др. Казань: Казан. гос. ун-т, 1994. С. 77–106.

Сулейманов Д.Ш. К вопросу о числе татарских падежей // Исследования в лингвистике / науч. ред. Р.Г. Бухараев и др. Казань: Фэн, 1996. С. 70–84.

Сулейманов Д.Ш., Гатиатуллин А.Р. Структурно-функциональная компьютерная модель татарских морфем. Казань: Фэн, 2003. 220 с.

Татарская грамматика / ред. М.З. Закиев, Ф.А. Ганиев, К.З. Зиннатуллина. Казань: Татар. кн. изд-во, 1993. Т. II. Морфология. 397 с.

Татарский национальный корпус «Туган тел» [Электронный ресурс] http://tugantel.tatar, дата обращения: 25.03.2017.

Хакимов Б.Э., Гильмуллин Р.А. К разработке системы параметров морфологической разметки для электронного корпуса татарских текстов // Труды Казанской школы-семинара по компьютерной и когнитивной лингвистике TEL–2008. Казань: Казан. гос. ун-т, 2009. С. 24–29.

Шимкова М. Репрезентативность корпуса как лингвистическая проблема // Труды международной конференции «MegaLing – 2005». СПб.: Изд-во «Осипов», 2005. С. 130–139.

Suleymanov D.Sh., Guilmoulline R.A., Guilmoulline A.A. Tatar phonological rules as a base of two-level morphological analyzer // Proceedings of LP'2000. Prague: The Karolinum Press, 2000. P. 495–504.

Сөләйманов Җәүдәт Шәүкәт улы,
техник фәннәр докторы, профессор,
Татарстан Республикасы Фәннәр академиясе академигы,
ТР ФА Гамәли семиотика институты директоры

Хәкимов Булат Эрнст улы,
филология фәннәре кандидаты,
Казан (Идел буе) федераль университеты Лев Толстой исемендәге
Филология һәм мәдәниятара багланышлар институтының
белем бирү технологияләре һәм филологиядә
мәгълүмати системалар кафедрасы доценты

Гыйльмуллин Ринат Абрек улы,
физика-математика фәннәре кандидаты,
ТР ФА Гамәли семиотика институтының
когнитив тикшеренүләр бүлеге мөдире