Татар теленең язма корпусы

Бүгенге тел белемендә 21-нче гасыр – корпуслар гасыры дип атала.

Нәрсә ул корпус?

Гомумиләштереп әйткәндә, корпус - ул күп сандагы текстлар җыелмасы нигезендә тәзелгән һәм компьютерлар хәтерендә электрон формада саклана торган сүзлек–белешмә. Корпуста, ике телле сүзлекләргә хас булган, сүзләрнең тәрҗемәсе дә, аңлатмалы сүзлекләргә хас булган төшендерүләр дә китерелми. Аның вазыйфасы телне мөмкин кадәр тулы һәм барлык үзенчәлекләре белән гәүдәләндерү.

Моңа чаклы телләр ике формада күзәтелеп килде. Бүгенге компьютерлар телне тагы бер формада – электроннар җирлегендә яшәү мөмкинлеге белән тәэмин иттеләр. Мәгълүм ки, бүгенге компьютерларның хәтере телне язма текстлар буларак та, сөйләм текстлар буларак та саклауга; белеме исә (программалар системасы) аларны бер формадан икенче формага күчерергә җитәрлек. Гадиләштереп әйткәндә, алар шәхес куйган сорауга сөйләм телендә дә, язма телдә дә җавап бирергә сәләтлеләр.

Шәхес куйган сораулар нинди булмаска мөмкин? Нәкъ менә шуның өчен дә тел компьютер хәтерендә яшәүче корпуста тулы һәм барлык үзенчәлекләре белән гәүдәләнергә тиеш була.

Бер уңайдан тел төшенчәсен дә ачыклап үтик. Бөекбритания энциклопедиясендә: “Тел – ул аерым халык, милләт тарафыннан фикер гәүдәләндерү, тапшыру максаты белән файдаланучы барлык сүз һәм сүзтезмәләрнең җыелмасы”- диелгән.

Электрон корпус шушындый җыелманы тәшкил итә дә инде. Милли корпус дәрәҗәсендәге корпусларларның күләме 100 миллион сүздән (сүзкулланыштан) ким булмый. (Инглиз теленең милли корпусы 400 млн. сүздән рус теленең милли корпусы 300 млн. сүздән тора).

Милли корпуслар язма тел, сөйләм тел, җирле һәм социаль дилектларны чагылдырырга тиеш булалар.

Корпусларга хас тагы бер сыйфат – ул, нинди типтагы лингвистик мәсьәләләрнең корпус ярдәмендә чишелү мөмкинлеген күздә тотып төзелә. Шул сәбәпле корпус, лингвистик мәгълүматлар белән генә түгел, күздә тотылган мәсьәләлрне чишү программалары белән дә тәэмин ителә.

Моннан тыш, корпуска кую алдыннан тексттагы язмаларга кайбер үзгәрешләр дә кертелә – тамгалар өстәлә (разметка ясала).

Әйтик, әгәр корпус ярдәмендә текстларны җөмләләргә бүлү күздә тотылса, ул вакытта җөмләләрнең чиген ниндидер тамгалар белән билгеләп чыгу кирәк булачак. Чөнки бары тик ноктага таянып кына компьютер җөмлә беткән урынны таба алмый.

Әгәр корпус ярдәмендә җөмләдә сүзләрнең нинди синтаксик функцияләр башкаруы өйрәнелсә, ул вакытта корпуста “агач көрәк” кебек сүзтезмәләрдә “агач” сүзенең сыйфат функциясе үтәвен күрсәткән

тамганың яки аны автоматик рәвештә ачыклый торган программаның булуы таләп ителә.

Кайсы типтагы лингвистик мәсьәләләрне чишүгә юнәлтелгән булуларына карап корпуслар аерым төрләргә бүленәләр:

1. Милли һәм милли булмаган корпуслар. Милли корпуслар күптөрле мәсьәләләр чишүгә мөмкинлек тудыралар һәм, алда әйтелгәнчә, күләм ягыннан 100 миллион сүздән ким булмыйлар. Милли корпусларга хас икенче бер сыйфат – алар репрезентативлык һәм бәрабәрлек (сбалансировланность) шартларына җавап бирергә тиеш. Репрезентативлык таләбенә җавап бирүче корпус бу телгә хас барлык үзенчәлекләрне дә чагылдырырга, ягъни текстлар җыелмасында сөйләм тел дә, , территориаль һәм социаль диалектлар да, телнең функциональ стиьләрен чагылдырган текстлар да урын табарга тиеш була. Бәрабәрлек шарты буенча һәр төр текст корпуска үзенең телдә күзәтелү күләменә бәйләнештә кертелә. Әгәр телдә публицистик стиьдәге тексталр башка төр стильләргә караганда күбрәк күләмдә күләмдә күзәтеләләр икән, димәк корпуста да алар күбрәк санда чагылырга тиешләр.

2. Нинди текстлар җыелмасын тәшкил итүләренә карап корпуслар язма (письменный), сөйләмә (устный) һәм катнаш (смешанный) булалар.

3. Телләр типологиясе, тәрҗемә (машина тәрҗемәсе) кебек мәсьәләләргә мөнәсәбәтле корпуслар бертелле (одноязычные), икетелле (двуязычные) һәм күптелле (многоязычные) булырга мөмкиннәр һ.б.

Әйтергә кирәк, корпусларга хас төрлелек әйтелгәннәр белән генә чикләнми. Аерым фәннәр телен, гасырлар дәвамында конкрет бер телнең үзгәреш динамикасын чагылдырган төрле чор корпусларының һәм хәтта ябык корпусларнң булуы да мәгълүм.

Татар теленең язма корпусына хас үзенчәлекләр.

Шушы елның апрель башында Татар теленең язма корпусы төзелеп Интернетка куелды. Исеменнән күренгәнчә, сөйләм теле корпуста юк. Шуның өчен аны милли корпус дип әйтеп булмый. Ә менә язма телебезне ул җитәрлек дәрәҗәдә тулы чагылдыра дип әйтә алабыз. Төзелгән язма корпусның күләме 45 миллионнан артыграк сүзгә тигез. Корпуска кергән сүзләр эчендә төрле сүзформаларның саны 400 мең чамасында. Корпус, татар әдәби телендә киң чагылыш тапкан, публицистика, матур әдәбият һәм гуманитар фән стилендәге текстлар җыелмасыннан тора. Корпуста публицистикага караган текстлар якынча 60%, матур әдәбиятка караган текстлар - 35%, фәнни текстлар -5% тигез.

Күпме соң ул 45 миллион сүз?

Чагыштыру өчен бер мисал китереп үтәбез. Әдәбиятта урын тапкан мәгълүматләр буенча, Ф.М.Достоевский әсәрләренең тулы басмасында ике миллион чамасы сүз бар. Интернетка күз салсак, ул басманың 30 томнан яки 33 китаптан торуын күрербез. Димәк, төзелүче электрон корпус бүгенге хәлендә дә татар теленә хас сүзләр һәм формалар төрлелеген, тулыснча булмаса да, 75 - 80 процентка гәүдәләндерә дип әйтергә мөмкин.

Корпус, аның программалар системасы (корпус менеджеры) тел мәсьәләләре белән кызыксынган шәхескә күптөрле ярдәмнәр күрсәтергә сәләтле. Шул исәптән:

- шәхес кызыксынган сүзне табарга һәм аның татар телендә нинди ешлык белән кулланылуын чамаларга,
- табылган сүзформаның алдыннан һәм артыннан килүе мөмкин булган сүзләрнең (сул һәм уң контекстларын) ачыкларга, аларның кулланылу ешлыкларын чамаларга,
- табылган сүзформаның җөмләләрдә нинди сүзләр белән семантик мөнәсәбәткә керү мөмкинлеген (семантик контекст) ачыкларга,
- 50 мисал – җөмлә белән табылган сүзнең телдә ничек файдалануын бәян итәргә.

Хәзерге вакытта корпусның функциональ мөмкинлекләрен киңәйтү буенча эшләр башкарыла. 2-3 айдан корпусның күләме дә, мөмкинлекләре дә артыр дип өметләнәбез.

Татар теле электрон корпусы ярдәмендә игътибарлы шәхес бүгенге буын телендә теге яки бу сүзнең, сүзформаның барлыгы-юклыгын, булган очракта аларның нинди ешлык белән кулланылуын бик тиз ачыклый ала. Корпусның мөмкинлекләренә кагылышлы берничә мисал карап үтик.

1. *Корпус* сүзе (лингвистик термин буларак) татар телендә табыламы?

Корпустагы мәгълүматлар бу сүзнең 45 миллионнан артык сүзләр җыелмасында биш мәртәбә очыравын күрсәтәләр. Әмма бу очрауларда *корпус* сүзе техник һәм хәрби терминнар буларак файдаланыла. (Сул һәм уң контекстлар составына игътибар итегез). Димәк бүгенге буын теленә *корпус* сүзе “электрон корпус” мәгънәсендә, ягъни лингвистик термин буларак күзәтелми.

2. Корпусны актарган шәхес бәтенләй көтелмәгән хәлләргә дә юлыгырга мөмкин. Мәгълүм булганча, дин, беренче чиратта, харам ризыклардан, эчемлекләрдән; начар эшләрдән, гадәтләрдән саклануга юнәлтелән тәгълимат. Ә менә корпус мәгълүматлары буенча *харам* сүзе буын телендә 140 мәртәбә, ә *хәләл* сүзе 3432 мәртәбә файдаланылган. Безгә калса, *харам* сзе *хәләл* сүзенә караганда күберәк күзәтелергә тиеш иде.

3. 20 – 21 гасырлар чигендә дөнья күргән техник терминнары булдыруга, булганнарын активлаштыруга караган сүзлекләр, һичшиксез, белгечләр тарафыннан файдаланыла торгандыр, әмма халыкка таратыла торган текстларда аларның кулланышы сизелерлек түгел. Ахырысы галимнәр дөньясын уртаклаша, халыкка якынайта алмадык.

4. Корпуста табылган мәгълүматлар татар теленә хас булган кайбер терминнарның бүгенге язмаларда русча вариантта очравын дәлиллиләр. Мәсәлән, *ваба (wаба)* сүзе 29, ә аның тәрҗемәсен тәшкил иткән *холера* сүзе 52 мәртәбә файдаланылган. Шулай ук *бүсер* сүзе 42 мәртәбә, *грыжа* сүзе 5 мәртәбә кулланылган булып чыкты.

5. Һәр тел аерым халыкның тышкы һәм эчке дөньясына, кабилә кабул иткән кыйммәтлекләр системасына мөнәсәбәттә яратылган һәм шул мәдәниятнең яшәешен тәэмин итә.

Татар теленә хас сыйфатларның берсе – аның фигыль формаларына бай булуы. Шул формаларның кайберсе процесс белән эш башкаручы шәхес арасындагы бәйләнешне (мөнәсәбәтне) чагылдыра. Татар теле грамматикаларында ул “юнәлеш” дип, рус тел белемендә “залог” дип исемләнә. Бу бәйләнеш, нигездә, процесс өчен эш башкаручының ни дәрәҗәдә җаваплы булуын тасвирлый. Мәсәлән, *күрсәтү* фигыленең *күрсәт* формасы - эш башкаручыны бары тик эшкә өнди, процесс әле башланмаган. *Күрсәтә* формасы исә – бу эшкә җаваплы шәхеснең кем булуын, аның кем тарафыннан башкарылуын белдерә. *Күрсәтелә* формасы – үтәлүче эшкә җаваплы кешенең билгесез булуын, я булмаса аңлы рәвештә аның роле кимителүен тасвирлый. *Күрсәттерә* формасы -башкаручының бу эшне үз ирке белән түгел, ә блки кемнеңдер ихтияры белән үтәлүен белдерә. *Күрсәтешә* формасы – башкаручының кемгәдер ярдәм йөзеннән бу эшкә алынуын бәян итә. Санап үтелгән процесс белән эш башкаручы арасындагы бәйләнешләрнең кайсылары татар дөньясында актив күзәтелә икән?

Корпус җирлегендә без берничә күчемле фигыльләрдә шушы формаларның күзәтелү ешлыкларын ачыкладык. Алар таблицада китереләләр.

Таблица

өндәү	казы	кис	эч	төя	җый	яса	ю.к.к.	%
к.саны	91	96	768	47	61	124	1187	7,0%
төп	казый	кисә	эчә	төйи	җыя	ясый		
к.саны	693	393	1237	43	2206	6650	11222	66,5%
пассив	казыла	киселә	эчелә	төялә	җыела	ясала		
к.саны	91	97	39	13	2341	1712	4293	25,4%
йөкләтү	казыта	кистерә	эчтерә	төятә	җыйдыра	ясата		
к.саны	10	17	0	5	9	92	133	0,8%
бергәлек	казыша	кисешә	эчешә	төяшә	җыеша	ясаша		
к.саны	0	30	0	2	3	0	35	0,2%
җыенысы	885	603	2044	110	4620	8578	16870	

Искәрмә.

Таблицаның беренче баганасында кыскартылган формада процесс белән эш башкаручы арасындагы бәйләнеш төрләренең (юнәлеш формаларының) исмлеге, күзәтелү саны (к. саны); сигезенче һәм тугызынчы баганасында бөтен саннар һәм процентлар белән анализланган фигыльләрдә юнәлеш төрләренең күзәтелү күләме (ю.к.к.) тасвирлана.

Нәтиҗәләрдән күренгәнчә, иң еш файдаланылган хәбәрләү формасы – ул процесснең кайсы эш башкаручы тарафыннан үтәлүен чагылдырган структура. Икенче урында – эш башкаручының кем булуын белдерми торган хәбәрләү формасы. Бер караганда, бу фактлар берсен-берсе инкарь

итәләр сыман. Башкаручының кем булуын ачык белдерү дә, башкаручының кем булуын күрсәтмәскә тырышу да тел ияләре көнкүрешендә берчама чагылыш таба. Миңа калса, хикмәт тел ияләренең ике төргә - түрәләр һәм түрәтүгелләргә бүленешендә. Түрәләргә һәм аларга хезмәт итүче гәзит- журнал хезмәткәрләренә *төзелә, күзәтелә, сөрелә, җыела* дип сөйләү уңай булса, түрәтүгелләргә *төзи, күзәтә, сөрә, җыя* дип сөйләү аңлаешлырак һәм ышанычлырак.

Мисаллар: “Корпус нәрсәгә кирәк”- дигән сорауга җавап итеп китерелделәр. Таблицада күрсәтелгән нәтиҗәләрнең бәлки әле башка аңлатмалары да бардыр.

Еш кына телчеләрнең үзләре ижат иткән кагыйдәләргә, ясаган искәрмәләренә әдәби әсәләрдән мисаллар эзләп утыруларын күрергә туры килде. Һәр телченең үз темасы буенча мисаллар туплавы да, картотекалар төзүе дә күпләребезгә мәгълүм. Электрон корпус аларны мондый эшләрдән азат итәчәк.

Шундый ук ярдәмне электрон корпус татар телен ана тел яки чит тел буларак укытучыларга да күрсәтергә мөмкин. Корпустан файдаланып секунд аралыгында уытучылар үзләренә кирәкле мисалларны таба, укучылар исә үзләре күрәсе килгән сүзләрне сүзформаларны күрә, аларның кулланылу формалары, язылышы белән таныша алалар.

“Мондый ярдәмнәрне шәхесләргә сүзлекләрдә күрсәтә ала, нигә алар өчен корпус төзергә?” – диючеләр дә табылыр. Әмма сүзлекләр белән корпуслар арасында тигезлек юк һәм була да алмый. Беренчедән, корпус (компьютер белән берлектә) шәхес беләсе килгән сүзне секунд эчендә тәкъдим итә. Икенчедән, ул сүзләрне табигый җирлектә - контекстлар эргәсендә күрсәтә. Өченчедән, корпус: “Бу сүз телдә бар” – дип кенә әйтми, ул аның ничә тапкыр файдалануын да белдерә, ягъни бирелгән мәгълүматларны объективлык белән дә тәэмин итә.

Соңгы елларда тел белемендә корпуслардан файдаланып телне өйрәнүче яңа фән - корпуслы лингвистика (Корпусная лингвистика) барлыкка килде. Корпуслы лингвистика ул компьютер һәм корпуслар симбиозы гына түгел, ул лингвистика белән социолингвистиканы (тел төзелеше белән тел кулланышын) бергә ялгап өйрәнүче фән. Корпуста чагылыш тапкан нәтиҗәләр татар тел структурасыннан тыш телнең файдлануын – аерым буынның сөйләмен (тел дөнъясын) дә тасвирлыйлар. Гасыр буе әдипләр телендә табылган аерым мисалларга нигезләнеп, я булмаса гәзит битләрендә күзәтелгән терминнарга таянып телебез байый, үсә дип килдек. Баксак, телебез югалу хәлендә булып чыкты. Кулланыш даирәләре тарайган, фәнни терминнарын югалткан аралашу системасы булуы ачыкланды.

Корпус төзелү якын арада татар теленең ешлыклар сүзлеге, аның морфологик төзелешен өйрәнү өчен мөһим булган кире сүзлек, татар теленә хас стильләр үзенчәлеген, аерым әдипләр ижатының стиль

үзенчәлекләрен чагылдырган сүзлекләр эшләнүгә дә мөмкинлекләр тудыра.

Телчеләргә электрон формада сакланучы тел байлыгы һәм аны тиз арада күзәтеп чыгу юлы тәкъдим ителә. Русларның милли корпусы рус телен корпуслы лингвистика фәненә индергән кебек, татар теленең милли корпусы да якын киләчәктә татар тел белемен корпуслы лингвистика кысаларында өйрәнүгә китерер дип ышанасы килә.

Әйтергә кирәк, корпус техник ярдәмче генә түгел ул – татар тел дөньясы. Бүгенге көндә татар тел дөньясы нинди? Нигә бүген тел ияләре аңындагы төшенчәләр, тел төшенчәләр буларак, шәхес аңында гәүдәләнеш тапмый? Шәхесләр уй-фикерләрен туган телләрендә әйтә алмый газап чигәләр.

Бүген телебез Интернетка чыкты. Язма татар телен файдаланучыларның исемлеге бермә-бер киңәйде. Интернетта чыккан текстларның авторы язучылар, журналистлар гына түгел. Киресенчә, гуманитар булмаган фән вәкилләре. Бу, һичшиксез, уңай күренеш – татар теленең эшкә җигелүе. Шуның белән бергә, блоггерларга хас булганча, татарча язучылар үзләренә “ишетелгәнчә язу” принцибын якын итәләр. *Әдәбияты* сүзе корпуста **әдәбияты** формасында 6815 мәртәбә**, әдәбияте** формасында 94 мәртәбә күзәтелә. Орфография күзлегеннән хата язу 1,38% тәшкил итә. *Әдәбиятын* сүзенә игътибар итсәк, аның **әдәбиятын** формасында 1206, **әдәбиятен** формасында 36 мәртәбә очыравын кәрәбез. “Хаталык” бу очракта 3% җитә.

Телне боргычламый, сүзләрне ишетелгәнчәрәк әйтергә тырышу *мәгълүмат* сүзе җирлегендә дә аык чагыла. Корпуста **мәгълүмат** формасында ул 20422 мәртәбә, ә **мәгълүмәт** формасында 178 мәртәбә кабатлана. “Хаталык” күләме 0,87%. Шул ук вакытта **мәгълүматын** формасы 81, ә **мәгълүмәтен** формасы 2 мәртәбә кабатлана. Бу очракта “хаталык” 2,47% җитә. Мисаллар санын бермә-бер күбәйтергә мөмкин.

Ишетелгәнчә язу принцибы алдагы көндә активлашачак. Шуны күздә тотып, бәлки тел галимнәренә орфографияне камилләштерү чараларын күрү кирәктер. Аның фәнни булмавы, телне яңа мәгълүмати технологияләргә иңдерүдә бихисап авырлыклар тудыруы күпләргә таныш.

Соңгы елларда Машина тәрҗемәсе, Автоматик рәвештә әйтелгәнне таный торган системаларның бары тик корпуслар ярдәмендә генә чынга ашу мөмкинлеге тәмам ачыкланды.

Әйтик, машина тәрҗемәсендә система *үпкә* сүзенә юлыкты ди. Бу сүз - омоним. Омонимнар, кагыйдә буларак, контексттан чыгып чишеләләр. Әмма аерым очракларда чишелеш сул һәм уң контекстлар хәленнән килмәскә дә мөммкин. Андый очракларда, гадәттә, текстның тематик үзенчәлегенә ихтибар ителә. Тематик үзенчәлек *үпкә* сүзнең семантик контексттында чагылыш тапмый калмый. Кыскасы, әгәр тәрҗемә ителүче *үпкә* сүзенең семантик контекстты, корпусттагы *үпкә* сүзенең семантик

контексттына охшаш икән, ул чагында *үпкә* “легкие” мәгънәсендә тәрҗемә ителергә тиеш була.

Татар теленең төрле чорга караган корпусларын булдыру, аларны үзара чагыштырып телебез кичергән хәлләрне, аның үткән юлын, киләчәген төсмерләү бүгенге чорда хыялга сыймаган эшләр түгел. Секунд эчендә кирәкле сүзне, кирәкле форманы; галимнәр, укытучылар, укучылар сораган мисалларны табуга сәләтле корпус, алда күрсәтелгән мәсьәләләрнең көн тәртибенә куелышына, чишелешенә ышаныч уята.

Корпус телне тулы килеш (репрезентатив) формада күз алдына китерү мөмкинлекләре тудыра. “Микъдар сыйфатка әйләнә” (Количество переходит в качество) дигән закончалыкның инкарь ителгәне булмады. Димәк, татар теленең милли корпусын төзүгә, аннан файдалануга галимнәр бурычлы да.

Язма корпус Интернетка куелган. Аның адресы:

http://corpus.tatfolk.ru

Өстәмә сүз. Корпусны эшләү бүгенге көндә дәвам итә. Без аның күләмен 100 миллион сүзгә җиткерәчәкбез. Бу вакыт эчендә корпус функциональ планда да баетылачак, ягъни бүген чишелми торган кайбер мәсьәләләрнең ике–өч айдан чишү мөмкинлеге туар.

Хөрмәтле галимнәр. Форсаттан файдаланып Сезгә бер үтенечебезне белдерәбез.

Без Сездән татар телендә басылып чыккан фәнни хезмәтләргезнең, әсәрләрегезнең электрон версиясен җибәрүне үтенәбез. “Безнең хисапка баемакчы булалар”- дип уйламагыз. Без язма корпус төзеп бер тингә дә баемадык. Сез дә корпустан түләүсез файдаланачаксыз. Бары тик эшегездә аңа таянуыгыз хакында әйтергә кирәк булыр. Анысы инде гадәти әхлак.

Сез җибәргән электрон версиягә килгәндә, тагы шуны белдерәбез. Әсәрләр корпуска җөмләләргә бүленеп, ягъни текстлар буларак түгел, ә бәлки җөмләләр берәмлегендә кертеләләр. Һәр җөмләгә карата аның кайдан алынуы, авторы кем булуы хакында мәгълүматләр биреләчәк. Бәндәләргә Сезнең иҗат җимешләрен бөтен текст буларак күчереп алу мөмкин түгел.

Сайхунов Мансур,
Ибраһимов Тәүзих,
Сәлимҗанов Илнар.