Татар теленең язма корпусы

21 нче гасыр башында Интернет битләрендә, телчеләр телендә “корпус”, “милли корпус” кебек терминнар күзәтелә һәм ишетелә башладылар. 1990 елда инглиз теленең милли корпусы, соңнанрак нимес, француз, итальян телләренең милли корпуслары төзелеп дөньяга таралды. 2007 елларда рус теленең милли корпусы төзелеп бетүе мәгълүм булды. Шул вакытлардан бирле төрле илләрдә милли корпуслар туа торды. Нәтиҗәдә компьютер һәм корпуслар ярдәмендә тел өйрәнүче фән тармагы “Корпуслы лингвистика” (Корпусная лингвистика) барлыкка килде, телчеләр арасында 21- нче гасырны милли корпуслар гасыры дип тану гадәткә керде.

Төрки телләрдә күрсәтелгән яңарышлар соңгы елларда гына күзгә чалына башладылар. Интернет хәбәр итүенчә, төрек теленең ике миллион сүздән торган тамгаланган (разметка ясалган) корпусы төзелгән. Кызганычка каршы, ул корпус белән танышу өчен регистрация үтү сорала (коммерческий корпус булырга охшый). Казах, кырым татарлары телләре өчен корпуслар булуы турында әйтелә, әмма аларның күләме, сыйфатлары хакында мәгълүматлар юк.

Нәрсә ул корпус?

Гомумиләштереп әйткәндә, корпус - ул күп сандагы текстлар җыелмасы нигезендә тәзелгән һәм 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 электрон формада 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 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 төр 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.

Кайсы төрдәге лингвистик мәсьәләләрне чишүгә юнәлтелгән булуларына карап корпуслар күп кенә типларга бүленәләр:

1. Милли һәм милли булмаган корпуслар. Милли корпуслар күп төрле мәсьәләләр чишүгә мөмкинлек тудыралар һәм, алда әйтелгәнчә, күләм ягыннан 100 миллион сүздән ким булмыйлар. Милли корпусларга хас икенче бер сыйфат - ул репрезентатив һәм сбалансировланган булырга тиеш.

2. Нинди текстлар җыелмасын тәшкил итүләренә карап корпуслар язма (письменный), сөйләм (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.

Без төзегән язма корпуска нинди сыйфатлар хас?

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. 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.

Корпусның программалар системасы (корпус менеджеры) тел мәсьәләләре белән кызыксынган шәхескә күп төрле мәгълүматлар бирергә мөмкин. Мисал өчен:

1. Шәхес кызыксынган сүзне табарга һәм аның татар телендә нинди ешлык белән кулланылуын күрсәтергә,

2. Табылган сүзнең алдыннан һәм артыннан килүе мөмкин булган сүзләрне (сул һәм уң контекстларны) ачыкларга һәм аларның кулланылу ешлыкларын бәян итәргә,

3. Табылган сүзнең төрле җөмләләрдә нинди сүзләр белән семантик мөнәсәбәткә керү мөмкинлеген (семантик контекст) күрсәтергә,

4. Табылган сүзнең телдә ничек файдалануын дәлилләп 50 мисал – җөмлә белән дәлилләргә.

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

Корпуслар төзү, аерым алганда татар теленең язма корпусын төзү нигә кирәк?

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 сфера һәм средасы тарайган, фәнни терминнарын югалткан аралашу системасы булуы ачыкланды.

Корпусның мөмкинлекләрен күрсәтү йөзеннән 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. Бүгенге буын язма татар телендә *әманәт* сүзе 158 мәртәбә, акция сүзе 2474 мәртәбә очыраган булып чыкты.

3. Корпуста чагылыш тапкан мәгълүматлар көтелмәгән дә булырга мөмкин. Безнең өчен түбәндәге зурлыклар көтелмәгән булып чыктылар. Корпус нәтиҗәләре буенча *харам* сүзе буын телендә 140 мәртәбә, ә *хәләл* сүзе 3432 мәртәбә файдаланылган. Безгә калса, *харам* сзе *хәләл* сүзенә караганда күберәк күзәтелергә тиеш иде. Чөнки дин, беренче чиратта, харам ризыклардан саклануга юнәлтелән тәгълимат.

4. Кызганычка каршы, корпуста 20 – 21 гасырлар чигендә дөнья күргән техник терминнары булдыруга, булганнарын активлаштыруга караган сүзлекләрнең телгә тәэсире сизелерлек түгел.

5. Корпуста табылган мәгълүматлар татар теленә хас булган кайбер терминнарның бүгенге язмаларда русча вариантта очыравын 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.

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.

Еш кына телчеләрнең үзләре иҗат иткән кагыйдәләргә, ясаган искәрмәләренә әдәби әсәләрдән мисаллар эзләп утыруларын күргә 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.

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

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 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.

Мансур Сайхунов,
Тәүзих Ибраһимов,
Илнар Сәлимҗанов.