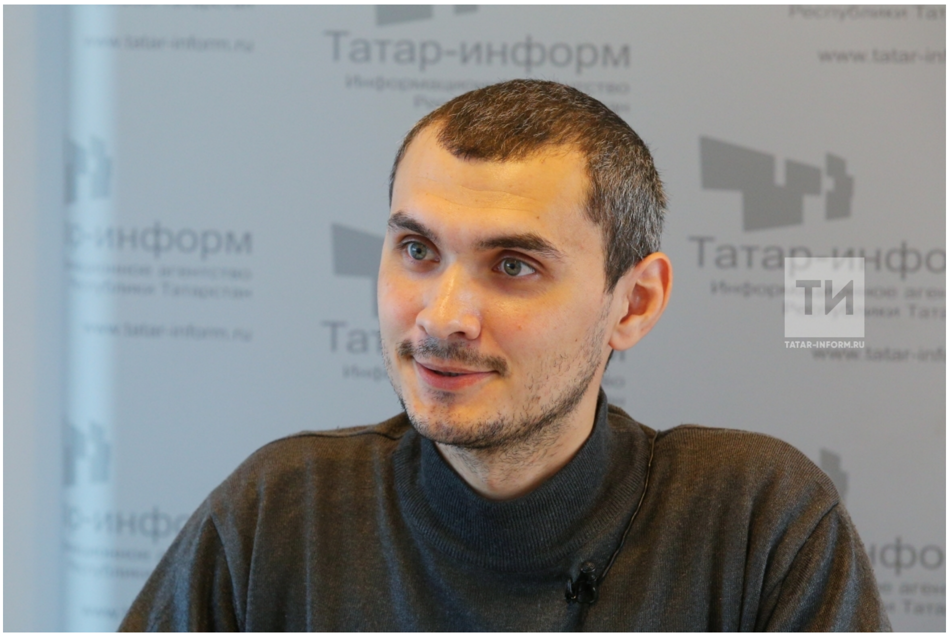
Татар дөньясы

Мансур Сайхунов: Татар теленең язма корпусы буш вакытыбызда үз акчабызга эшләнде

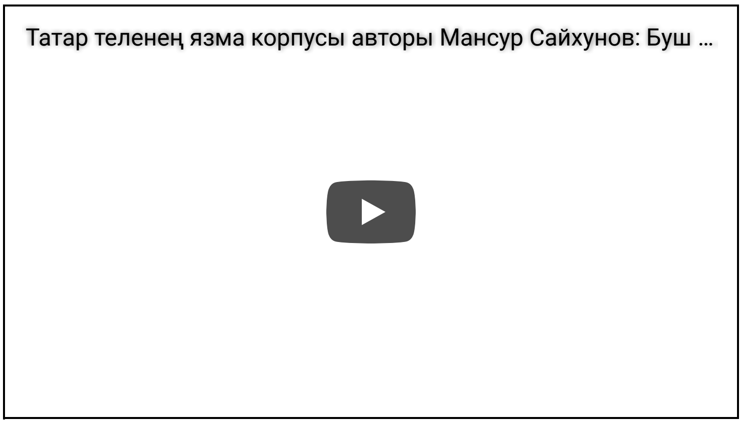
16 июнь 2017, 09:29 , Фото: Султан Исхаков

Татар теленең язма корпусы авторларының берсе фикеренчә, Интернет өлкәсендә яңа төрле проектлар эшләү өчен тел өлкәсендә эшләүчеләрнең берләшүе кирәк.

(Казан, 16 июнь, "Татар-информ", Рәмис Латыйпов). Татар теленең язма корпусы проекты - тулысынча иҗтимагый башлангычта эшләнгән проект. Проект авторларының берсе Мансур Сайхунов “Татар-информ” агентлыгына интервьюсында ни сәбәпле бу сайтны эшләүләрен, корпусның тел галимнәре, журналистлар, редакторларга ни дәрәҗәдә файдалы булуын, ни сәбәпле бу проектны үз акчаларына эшләүләрен сөйләде.



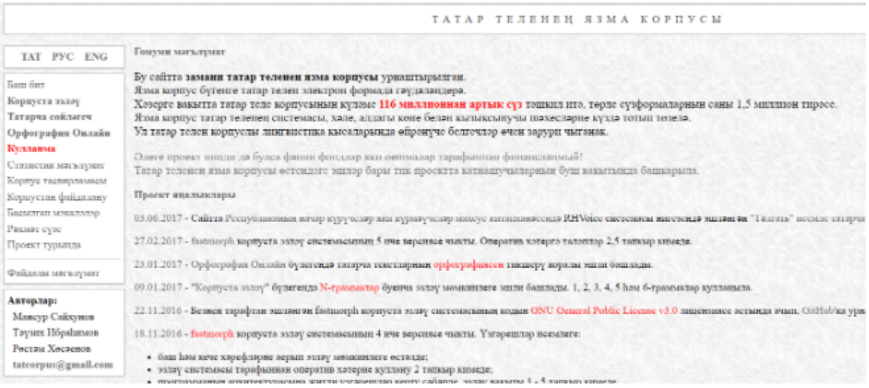
- Белешмә. Тел корпусы – текстларны электрон формада туплап, алардагы закончалыкларны күрсәтә ала торган мәгълүмати-белешмә системасы. Татар теленең язма корпусы - шундый проектларның берсе.



Корпуска нигез салыну: 2010-2011 еллар

- Мансур, татар теленең язма корпусы проектына кайчан керештегез, бу эшкә тотынуыгызның сәбәбе нинди?

- Университетта аспиратураны тәмамлап, кандидатлык диссератациясен яклагач, фән өлкәсендә бераз яңачарак юнәлешкә кереп китәсе килгән иде. Минем аңа кадәр эшләгән юнәлешем шактый кызыклы - сөйләм синтезлау өлкәсенә карый иде. Ләкин ул юнәлештә эшләр сүлпәнәйде. Без анда бер команда булып эшли идек. Эшләр сүлпәнәю аркасында башка юнәлеш турында уйлана башладык. Фәнни җитәкчем дә бу өлкәдә кызыксынучы кеше һәм башка килде инде: “Нәрсә булыр икән башлап карасак”, - дигән уй. Чөнки аны бит эшләмичә, берничек тә белеп булмый, барып чыгамы ул юкмы.

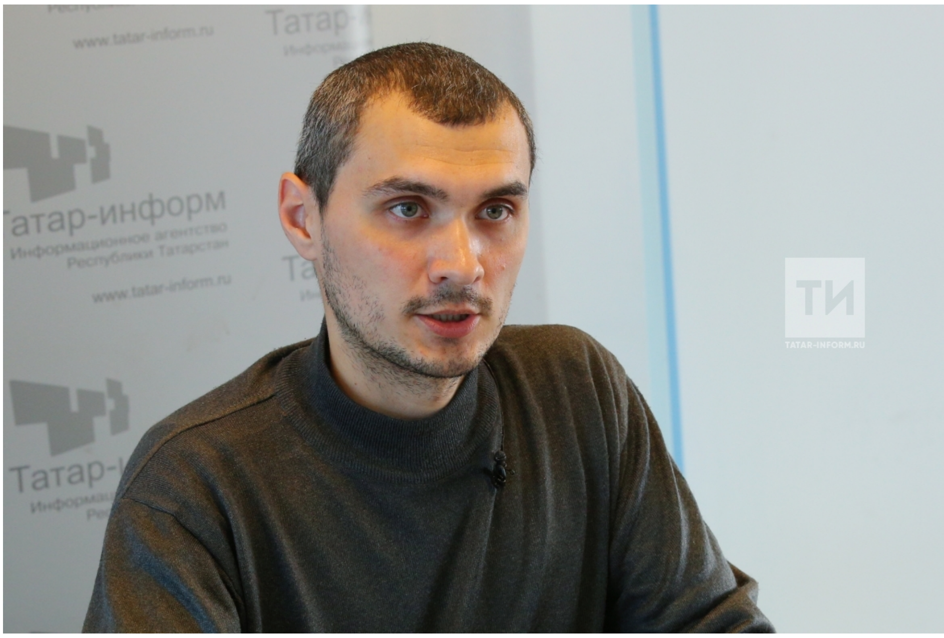


Татар теленең язма корпусы

Бу 2010-2011 еллар иде, күп киңәшләшүләр, сөйләшүләр булды. Акрын гына шулай башлап киттек, эшләр бара, белмәгәнен өйрәнәсең. Чөнки бу бөтенләй яңа нәрсә түгел, дөнья фәнендә - инглиз, испан, француз, рус телләрендә бу өлкәдә шактый эшләр башкарылган. Без юл ярып баручылар түгел идек, башкалар эшләгән тәҗрибәне үзебезнең туган телебездә кулланырга гына иде безнең планыбыз.

Безнең эш 2010-11 елларда башланды, 2012 елда беренче версиясе чыкты, без аны интернетка урнаштырдык. Аны беренче чиратта фән өлкәсендә эшләүчеләргә дип планлаштырган идек, шуларны күз алдында тотып төзелгән иде бу. Тел, әдәбият һәм сәнгать институты директоры Ким Мөгаллимовичка рәхмәт, ул институтта презентация ясарга шартлар тудырды. Шактый гына кешеләр җыелган иде. Шуннан соң проект эшләп китте, кулланучылар үзләре дә тәкъдимнәрен әйтә, нәрсәдер сорый башладылар.
Шулай итеп, корпус буенча эшләр берничә юнәлештә барды. Беренчедән, корпусның күләмен зурайтырга кирәк.

Корпус нәрсә ул? Башта шуны әйтеп китик әле. Корпус ул билгеле бер эзләү системасы белән җиһазландырылган зур сандагы текстлар җыелмасы. Бу текстлар җыелмасы телне мөмкин кадәр тулырак итеп чагылдырырга тиеш. Фән хезмәткәреме, башкалармы, телдәге ниндидер күренешләрне эзләр өчен корпуска керә ала. Күренешләр төрле өлкәдән булырга мөмкин, шуңа күрә корпуста фән, гуманитар, техник, матур әдәбият, публицистика текстлары да билгеле бер пропорциядә булырга тиеш.



Бу эшләрне башкаруда иң авыры текстлар җыю булды. Шактый гына текстларны интернеттан җыйдык. Бу проект турында белгән кешеләр үзләре дә мөрәҗәгать иттеләр, хәтта кайбер язучылар да: “Минем китапларны сезгә бирсәм, аласызмы?” - дип сораштылар. Ничек инде алмыйбыз, бик зур рәхмәт әйтеп алабыз. Кайбер текстларны газета-журнал чыгаручылар бирде, мәсәлән, Шакирҗанов Рәшит абый. Ул Фәрит белән безгә бик зур ярдәм итте. Алар күп еллар буе “Фән һәм тел” журналын чыгарып килделәр, анда нәкъ менә фәнни темага багышланган мәкаләләр бик күп иде, аларның ярдәме бик зур булды. Казанда фольклор үзәге бар, аның директоры Фәнзилә Җәүһәрова үзләренең конференция җыентыкларын бирде.

Матур әдәбиятка килгәндә, язучыларның биргән әсәрләреннән тыш, шактый гына интернеттан да тупладык. Туплаудан тыш параллель рәвештә корпусның функциональ ягын да үстерү кирәк иде. Эзләү өчен күбрәк мөмкинлекләр булсын өчен 2014 елларда без халыкара Apertium проекты эшләгән системаны кулланып, автоматик-морфологик разметка ясадык, ягъни корпуста бөтен сүзләр дә сүз төркеме ягыннан да, грамматик категорияләр ягыннан да тамгаланган һәм алар буенча да эзләргә була. Сүз уңаеннан ул проектта Илнар исемле татар егете дә катнаша.

Корпуста бүгенге көндә сүзформа буенча, шаблон буенча (йолдыз билгесе яки сорау билгесе куеп), грамматик категорияләр буенча, аннары сүзләрнең үзара арасы буенча эзләргә була.

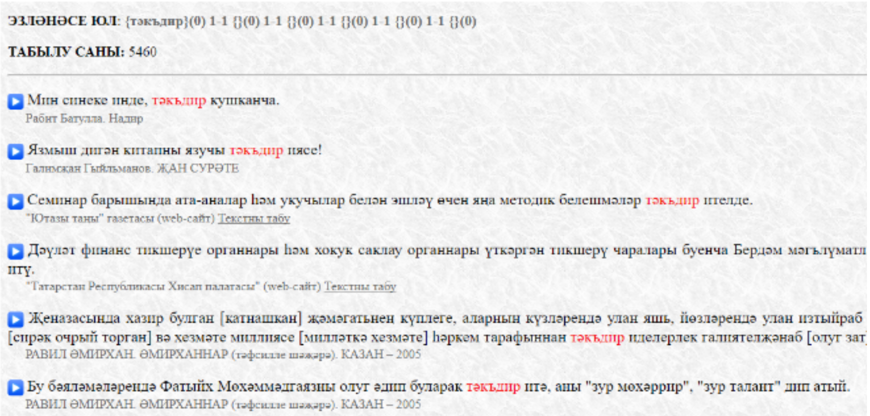
“Корпус - телдәге күренешләрне күрсәтүче система”

- Мансур тагын аңлатып китсәгез иде кулланучыларга, бу сайтта нәрсәләр эшләргә була? Сайтта бит сүзнең дөрес язылышын тикшерергә дә, кайсы сүзләрнең ешрак кулланылганын да белергә мөмкин.

- Беренчедән, сайт берничә категориядәге кешеләргә файдалы булыр дип уйлыйм. Иң төп аудитория - фән өлкәсендә эзләнүчеләр. Грамматик категорияләр телдәге ниндидер күренешләрнең булу-булмавын дәлилләү булып тора. Әгәр телдә ниндидер очраклар кулланыла икән, аларны күрергә була. Әгәр бу очрак башка очракларга караганда күбрәк яки азрак кулланыла икән, аны шундук тиз таба ала ул. Кешегә меңләгән китаплар карап эзләп утырырга кирәкми, корпус аны берничә секунд эчендә күрсәтеп бирә. Бер секунд, ярты секунд инде ул.

- Мәсәлән, ничек инде ул, мисал китерсәгез?

- Мәсәлән, шул ук “тәкъдир итү” сүзтезмәсе ничек кулланыла? Нинди́дер искерәк яисә гарәп теленнән кергән һәм бүгенге көндә берничә төрле язылу ихтималы булган сүзләрнең дөрес язылышы ничек? Кайсы сүз төркеме башка сүз төркеме белән килә ала? Гомумән, шуларны карап була.


ЭЗЛӘНӘСЕ ЮЛ: {тәкъдир}(0) 1-1 {}(0) 1-1 {}(0) 1-1 {}(0) 1-1 {}(0) 1-1 {}(0)

ТАБЫЛУ САНЫ: 5460

Мин синеке инде, тәкъдир кушканча.
Рабит Батулла. Надир

Язмыш дигән китапны язучы тәкъдир иясе!
Галимҗан Гыйльманов. ҖАН СУРӘТЕ

Семинар барышында ата-аналар һәм укучылар белән эшләү өчен яңа методик белешмәләр тәкъдир ителде.
"Ютазы таңы" газетасы (web-сайт) Текстны табу

Дәүләт финанс тикшерүе органнары һәм хокук саклау органнары үткәргән тикшерү чаралары буенча Бердәм мәгълүматл итү.
"Татарстан Республикасы Хисап палатасы" (web-сайт) Текстны табу

Җеназасында хазир булган [катнашкан] җәмәгатьнең күплеге, аларның күзләрендә улан яшь, йөзләрендә улан изтыйраб [сирәк очрый торган] вә хезмәте миллиясе [милләткә хезмәте] һәркем тарафыннан тәкъдир иделерлек галиятелҗәнаб [олуг зат
РАВИЛ ӘМИРХАН. ӘМИРХАННАР (тәфсилле шәҗәрә). КАЗАН – 2005

Бу бәяләмәләрендә Фатыйх Мөхәммәдгаязны олуг әдип буларак тәкъдир итә, аны "зур мөхәррир", "зур талант" дип атый.
РАВИЛ ӘМИРХАН. ӘМИРХАННАР (тәфсилле шәҗәрә). КАЗАН – 2005

Мисал. "Тәкъдир" сүзенең нинди сүзләр белән янәшә килүен ачыклау

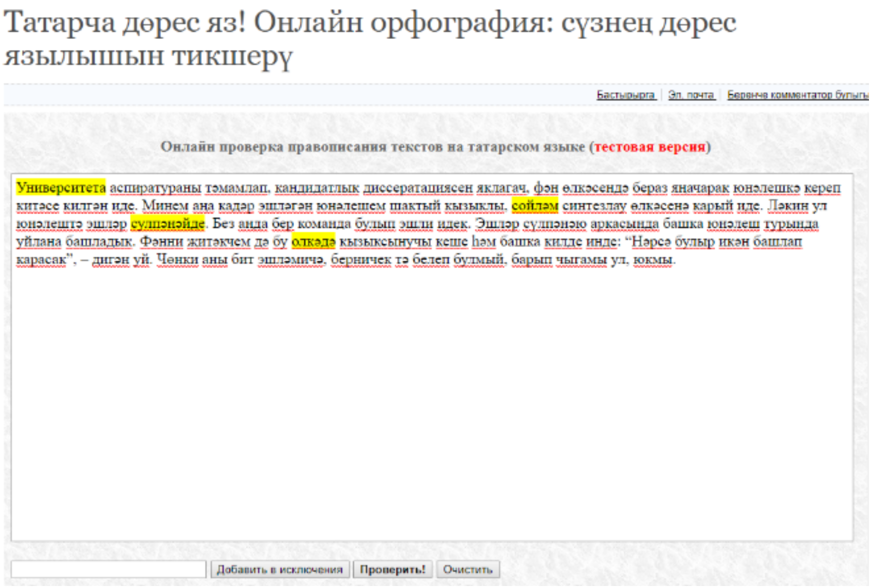
Бу корпус фәнни эзләнүләр өчен корал гына булып тора. Башыңа килгән идеяләрне тиз тормышка ашыру өчен корал ул.

Әле беркөнне генә профессор Камил Рәхимович белән сөйләшү вакытында кызык кына нәрсә әйтте. Танышы аңа бер китап күрсәткән. Моңа мисалларны шактый озак еллар эзләдем, тупладым, монда катлаулы мисаллар, дигән. Камил Рәхимович аңа: “Менә мондый сайтны кулланмадыңмы, монда бит бу мисалларны сиңа берничә секунд эчендә табып биреп була”, - дигән. Теге кеше, Камил Рәхимович күрсәткәч, шаккаткан инде, дөресен генә әйткәндә. Ягъни бу сайт вакытка ике-өч кенә түгел, меңләгән тапкыр экономия ясый. Һәм кеше бит китап укып утырганда да бөтен мисалларны күреп бетермәскә мөмкин, алар күзгә чалынмаска да мөмкин, ә бу бөтенесен табып бирә. Бу бер категория.

Икенче категория – корпус төрле нәшриятларда эшләүче редакторлар өчен файдалы булырга тиеш. Чөнки редакторлар нинди́дер текстны төзәтәләр, дөреслеген тикшерәләр икән һәм анда кеше язган нинди́дер сүзтезмә яки җөмләнең конструкциясе шиклерәк, мондый очрак телдә кулланылмый кебек тоела икән, ул аны корпуста тикшереп карый ала. Шунда язып карарга була - башка кулланучылар бу сүзтезмәләрне шулай бергә кулланаламы? Шундый тикшерүләр үткәреп була.
“Корпус хаталарны да тикшерә ала”

Корпусның тагын бер функциясе - татар текстларындагы орфографик хаталарны тикшерү

Шул корпус нигезендә орфография тикшерү системасын да эшли башладык. Ул система әлегә камил түгел, анда барлык хаталар да табылмаска яки кайбер дөрес сүзләр дә хаталы итеп күрсәтелергә мөмкин. Ләкин әле бу эш башланды гына, алга таба ул камилләштереләчәк. Редакторларга да текстларны, вак, механик хаталарны – “опечатка” дибез инде, белмәүдән түгел, ә ялгыш кына башка хәреф куелып киткән очракларны да бик тиз табу өчен кирәкле әйбер инде бу.



Мисал. Хаталы текстны тикшерү үрнәге. Татар теленең язма корпусының проекты "Интертат" сайтында

Аның идеясе дә үзем ниндидер мәкалә язып утырганда башка килде. Яңадан укып чыгасың - бер хата таптың, тагын бер укып чыктың - тагын бер хата. Беренче тапкырда ук табылмый бит барысы да. Шуңа берәр нинди система эшләргә була бит инде. Корпус ул башка бик күп проектлар өчен нигез булып тора һәм орфография тикшерү системасы шуларның берсе.

Корпусны кулланучыларның өченче категориясе дип компьютер лингвистикасы өлкәсендәге башка проектларны атарга була. Мәсәлән, сөйләмне тану системалары. Сөйләмне анализлау, “распознавание речи” була русча. Мәсәлән, кеше нәрсәдер сөйли ди, компьютер аны танып, язма формага күчереп барырга тиеш. Кеше бөтен авазларны да яңгыратып бетерми, сөйләмдә сүзнең бер өлеше йотылып кала. Мәсәлән, Казан дигәндә дә [а] авазы яисә кайбер тартыгы тулысынча яңгырамый. Озынрак сүз булган саен, йотылу очраклары тагын да күчләрәк була. Шуңа күрә система авазны ишетеп кенә беркайчан да төгәл сүзне танып бетерә алмаячак. Кеше “Казан дәүләт университеты” дип әйтте һәм сүзнең азагы, озын җөмлә булу сәбәпле, йотылып калды, начар яңгырады ди. Компьютер системасы “казан”, “дәүләт”не таныды, ди, ә менә “университет”ның ахыры йотылып калды. Һәм нишләргә? Ул аны “уни” дип кенә калдырса, дөрес түгел инде ул. Бу очракта система фараз кыла - нәрсә булырга мөмкин? Монда корпусның статистикасы ярдәм итә, “Казан дәүләт” дигән сүзтезмәдән соң гадәттә нәрсә килә? Куллану ешлыгы буенча “Казан дәүләт университеты” татар телендә еш телгә алына. “Казан дәүләт универмагы” булуы бик шикле, андый сүзтезмәне кулланганыбыз юк. Бу корпус ярдәмендә сөйләмне тану системасының дөрес вариантны сайлап алып, матур гына җөмлә төзү ихтималы арта.

Бу сөйләмгә генә кагылмый, pdf форматындагы документларны, китапларны сканерлап, текстка әйләндерүче FineReader кебек программалар да бар бит. Китап таушалган булу сәбәпле, анда кайбер сүзләрдә хәрефләр бозылган булырга мөмкин. Шулай ук сүзнең бер өлеше югалып калса, корпус анда да ярдәм итә алачак.

Мондый тел корпуслары фәндәге башка бик күп проектлар өчен нигез булып тора, аларның кыйммәте дә шунда. Корпус башка өлкәләрдәге эшне күпкә җиңеләйтә. Корпуслар алар бер төрле генә түгел, бик күп төрле корпуслар бар, мәсәлән, параллель. Берничә телле корпуслар да бар, мәсәлән русча-инглизчә. Алар тәрҗемәчеләр эшендә бик ярдәм итә, чөнки менә бу конструкция ничек тәрҗемә ителә икән дип карап була.

Безнең корпус бер телле корпус, ул татар телен тулысынча чагылдыру максатын куя һәм шул юнәлештә эзләнүләр өчен бер нигез булып тора. Текстларны эзләү системасы бар. Төрле эзләнүләр өчен төрле функцияләр кирәк бит инде, һәм икенче яктан корпус эшен тиз башкарырга тиеш. Бернинди фәнни эзләнүче дә эзләү төймәсенә басып, тагын бер сәгать эзләүне көтеп утыра алмый инде, ул аның продуктивлыгын киметә. Әмма хәтта алай булса да, кул белән эзләүгә караганда күпкә тизрәк булыр иде.

- Орфографик тикшерүләрне ни рәвешле камилләштерергә мөмкин?

- Беренчедән, безнең планда сүзлекләр куллануны киңәйтү. Берничә сүзлек кулланыла инде, шуларны тулыландыру. Икенчедән, төп эш - Apertium проектының морфологик анализаторы белән корпусны ныграк тоташтыру. Чөнки Apertium проектының шул морфологик анализаторы нигезендә орфография тикшерү системасы да эшләнә һәм шул системаларның бөтенесен бергә берләштерә алсак, моның хаталар табу сыйфаты бик югары булачак.

“Корпуста 116 миллион сүз бар”

- Корпуска барлыгы күпме текст кертелде?

- Корпуслар, гадәттә, күләм буенча үлчәнә. Безнең корпусның беренче версиясе 45 миллион сүздән тора иде. 2014 елда чыккан икенче версия 116 миллион сүзне үз эченә ала, ягъни шуның кадәр күләмдәге текстлар кергән.

Текстлар санына килгәндә, бу бик субъектив күренеш. Текстларны ничек бүләсең бит әле. Мәсәлән, бездә матур әдәбият текстлары кайбер очракта аерым әсәр буларак, кайчак бер китап буларак берләштереп бирелә. Шуңа күрә аларның санын ничек исәплисең? Газета-журналларга килгәндә, шундый ук хәл. Без, гадәттә, аларны я санлап бирәбез, мәсәлән, бер санга кергән бөтен текстлар бер берәмлек булып тора. Интернеттан алынган ресурс булса, бер интернет-ресурсны бер чыганак итеп күрсәтәбез.

Ник аерым-аерым эшләмисез, алай җайлырак булыр иде дигән сорау бирсәгез, моңа җавап инде бик гади, чөнки ул вакыт таләп итә, аңа бик күп кеше ресурслары кирәк, аны төгәл билгеләп утырырга кирәк. Билгеле, акрын гына ул эш тә бара. Ләкин безнең проект ул ниндидер фондлар яки грантлар, оешмалар тарафыннан финансланмый. Бу проект безнең хобби буларак башкарыла. Командабызда берничә кеше бар - мин, фәнни җитәкчем Тәүзих абый Ибраһимов, тагын дустыбыз Рөстәм Хөсәенов. Команданың үзәге шулар һәм аңа еллар дәвамында кушылган башка кешеләр бар, мин аларны консультантлар дип атыйм. Чөнки телче булып чыктың икән университеттан һәм син корпус төзи аласың дигән сүз түгел. Монда техник белем кирәк. Беренчедән, программалаштыру өлкәсе, икенчедән компьютер системаларын администрацияләү белеме дә кирәк, чөнки корпус ул теләсә нинди веб-хостингка урнаштырдың да эшли ала торган система түгел, аңа махсус көйләнгән компьютер булырга тиеш. Аллага шөкер, минем IT бүлеге мөдире булып эшләвем дә ярдәм иткәндер, чөнки ул белемнең монда да кирәге чыкты. Төрле шартларның туры килүе бәхеттер инде ул. Әгәр дә ниндидер белем җитми икән, шул консультант дусларыбызга мөрәҗәгать итәбез, Аллага шөкер, барысы да булышып торалар һәм аларның саны арта бара.

“Интернетта татар телендәге текстлар бик күп”

- Бу корпусны эшләгәндә татар теленә карата нинди фикерләр туды, консультантлар нәрсәләр әйттеләр? Безнең теленең нинди үзенчәлекләре бар?

- Әйдәгез, мин телнең үзенә кагылмыйм. Бу өлкәдә эшли башлаганда иң беренче туган фикерләр - татар телендә мондый ресурсларның бөтенләй эшләнмәве иде. Ресурслар күбрәк булган саен яхшырак. Мәсәлән, монда корпус ди, тегендә параллель рәвештә морфологик анализатор эшләнеп килде ди. Ниндидер башка системалар, шул ук синтаксик анализатор булса, аларны бер-берсенә берләштереп, тагын да катлаулырак, тагын да файдалырак системалар эшләп чыгарып булыр иде. Башка ярдәмче системаларның булмавы, аларның бик аз булуы дип әйтик, эшне бик нык авырайта, тоткарлый.

Ә татар теленә килгәндә, нәрсә дип әйтим инде? Баштагы этапта мине шаккаттырган бер нәрсә - интернетта татарча текстлар алай күп түгелдер дип уйлый идем. Эзли торгач, бик күп ресурслар табылды, Аллага шөкер. Шәхси блоглар да күп икән бездә. Кешеләр үзләренең тормышы турында, ниндидер сәяхәтләре турында блоглар алып баралар. Мин укытучыларның блоглары турында әйтмим дә инде, аларныкы бик күп, алар тиештер дә, бәлки, аларны алып барырга. Шәхси блоглар да шактый. Анда бит укытучыларның үз лексикасы кулланыла, аларның үз стиле һәм алар да бик зур байлык инде безнең өчен. Татар теленең интернетта шактый зур колач алуы шаккаттырды һәм бик сөендерде.

“... әмма тенденция кимүгә таба бара”

Бүгенге көндә текстлар кимүгә таба охшаган. Ресурслар кими бүген. Менә мин корпуска тагын нинди текстлар кертергә була икән дип караштыргалыйм. Элек ачылган күп кенә ресурслар ябылды. Белмим, нинди сәбәптәндер. Бәлки, кешенең кызыксынуы беткәндер, бәлки, акчасы. Чөнки хостинг өчен дә түләп торырга кирәк. Бәлки кешенең вакыты юктыр.

Минем күзәтүләрем буенча, ресурслар кими бара. Кайбер яңалары чыга, әмма, гомумән алганда, кими кебек тоелды.

- Ә сез бу корпусның кайсы өлкәсен үстерергә җыенасыз, тагын нинди ресурслар кертеләчәк, нинди куллану очраклары булачак?

- Безгә төрле нәшриятлар текстларын бирсәләр, бик әйбәт булыр иде. Аларда бик күп китаплар басылып чыга бит. Безгә китапларны биргән очракта бернинди авторлык хокуклары бозылмый, чөнки корпуска кергән текстлар шундук җөмләләргә бүленеп, төрле рәвештә бутала. Корпусның системасын, русча әйтсәк, взлом ясап, андагы базаны урлый калсалар да, хәтта шулай булса да, ул текстларның бөтенлеген кире төзи алмыйлар. Чөнки җөмләләр буталган. Уйлап карагыз - корпуста бүгенге көндә 10 миллионнан артык җөмлә кергән текстлар. Аларны кире җыю мөмкин дә түгел, шуңа күрә аның авторлар өчен бернинди куркынычы юк, бернинди хокуклар да бозылмаячак. Ләкин эзләнүчеләр, шул ук нәшриятларда эшләүче редакторлар өчен аның файдасы бермә-бер артачак.

Алга таба бу корпустагы орфография тикшерү системасын камилләштерәсе иде. Аннары минем фәнни җитәкчем башлап җибәргән проект инде ул, татар сөйләмен синтезлый торган проектны (сүз «Тәүзих» татарча сөйләм җыйнагычы турында бара) яхшыртасы иде. Ул текст формасында язган материалны сөйләмә формага күчерә. Элек, университетта укыганда фәнни җитәкчеләрнең зур төркеме бар иде, шул вакытта мине чакырдылар, компьютер лингвистикасы белән кызыксынып киттем, әмма, кызганыч ки, ул проект акрынлап сүлпәнләнде. Без берничә ел элек шул проектны торгызып интернетка урнаштыру мәсьәләсен үз алдыбызга куйдык. Ләкин ул компьютерга урнаштырыла торган программа иде, интернетта эшләми. Шуңа күрә аның күп кодларын яңадан язарга туры килде, әле хәзер дә ул язылып бетмәгән, ләкин, гомумән алганда, эшли. Безнең корпуста кеше татарча текстлар язып, яңгыратып карый ала.

Мәсәлән, кухняда нәрсәдер эшләгәндә әдәби әсәрләрне тыңларга була. Әлбәттә, аның яңгырашы дикторныкы кебек матур, чиста түгел, кайбер төгәлсезлекләр дә бар, ләкин аңларга мөмкин. Кешегә бу файдалы була ала.

«corpus.tatar домены туган көнемә бүләк булды"

- Ни өчен сез .tatar доменында теркәлдегез?

- Бу безнең командадагы Рөстәм Хөсәеновның идеясе иде. "Ишеттеңме, тиздән татар домены булдырылачак, бу шундый мөмкинлек бит, татар теле өчен шундый проект эшлибез икән, әйдә шунда теркәлик", - диде. Уйлаштык та, онытылды да кебек, аннары, берзаман, безнең командада булган танышлар аны минем туган көнгә бүләк итеп эшләделәр. Шул доменны теркәп, акчасын түләп, миңа бүләк итеп эшләделәр. Ул чыннан да бик файдалы булды.

Аңа кадәр без бер дәүләт оешмасы доменында теркәлгән идек, ул оешманың директоры Фәнзилә Хәкимовнага бик зур рәхмәт. Башлангыч чорда ул мөмкин кадәр ярдәм итәргә тырышты.

"Татар доменында эзли торган Яндекс, Google кебек сайт кирәк"

- Тагын татар проблемасына әйләнеп кайтсак, татар домены ни дәрәҗәдә әһәмияткә ия? Без үзебез татар доменында эшләүчеләр буларак шуны сорыйсы килә – татар телендә сайтлар ачканда, татар доменында теркәлергә кирәкме? Нинди киңәш бирер идегез, татар доменының нинди өстенлекләре бар?

- Әгәр сайт татар теле өлкәсенә карый икән, минемчә, аны татар доменында теркәргә кирәк. Моның уңай ягы шунда - киләчәктә кешеләр нәкъ менә татар теле өлкәсендәге нindидер ресурсларны эзләсәләр, алар бер урында җыелып куелган булачак. Әйтәм бит, мин татар сайтларын эзләргә бик күп вакыт әрәм итәм. ru, com доменнарында да алар һәм алар арасында сайлап утырырга кирәк. Татар доменын кулланганда ул эш кимиячәк.

Иң мөһиме - шул татар доменында тирәнтен эзли торган система булырга тиеш, мәсәлән, Яндекс, Google кебек системалар татар доменында татар сайтларында эзләү мөмкинлеге бирерләр иде. Бу безнең ВМКдан чыгучы студентлар өчен, башка бер кызыксынучы белгечләр өченме бер идея да бит

инде. Шундый система эшләсәләр, кулланучылар табылачак. Чөнки татар сүзләрен эзләргә җиңел дип ничек кенә мактамасыннар, Яндекста да, Googleда да татарча эзләү системасы камил түгел. Озак эзләргә туры килә татарча, мин канәгать түгел. Яндекс бит, иң беренче чиратта, ru домены өчен эзли торган система буларак барлыкка килде һәм бүген дә, нигездә, шул өлкәдә лидер булып тора. Башка доменнардан эзләсә дә, беренче чиратта, ул ru домены өчен эшләнде. Татар домены өчен дә шундый берәр проект кирәк. Аның өчен әллә нәрсә кирәкми дә инде, минемчә. Чөнки ru доменында миллионлаган сораулар булса, татар доменында алай ук булмаячак. Шуңа күрә аңа әллә нинди кыйммәтле, көчле серверлар да кирәкми инде, ләкин кулланучылар табылачак. Аны эшләүчеләр алга таба реклама нигезендә дә акча эшли алырлар иде дип уйлыйм.

- Реклама, акча мәсьәләсенә килгәндә, Сез ничек тә булса акча эшли аласызмы?

- Акча эшләү бу эшкә керешкәндә үк бөтенләй күз алдында тормады. Шуңа күрә без акча эшлик, болай эшләп карыйк, дигән уйның башка да килгәне юк. Килмәстер дә инде. Кулланучылар бармы дигәндә, бәлки, эшләп куйганбыздыр да, ә чынлыкта кулланучы да юктыр әле, тиккә вакыт әрәм итәбездер. Бер дүрт ел элек сайтка счетчик куйдык. Корпуста эзлиләр микән дип беләсе килә иде, ул вакытка сайт ике-өч ел эшли иде. Кеше эзли башлагач, эзләү төймәсенә баскач, ул счетчик бергә арта бара. 3-4 ел элек ел буена 6,5 мең ярым тапкыр эзләү формасын тутырып сораучы булды. Ике ел элек бу күрсәткеч елына 12 меңгә кадәр үсте, былтыр 22 мең иде. Ягъни, шактый күп арта бара.

Ләкин эзләүчеләр саны быел әллә ни күп үсәр дип уйламыйм. Чөнки татар телен, инглиз, рус телләреннән аермалы буларак, күп кеше кулланмый. Куллану даирәсе аның кадәр чиксез түгел, ул өлкәдә эшләүче фән хезмәткәрләре дә бик күп түгел. Безнең татар теле өлкәсендә эшләүче хезмәткәрләр нигездә кайда? Казан федераль университетында, фәннәр академиясенең тел, әдәбият һәм сәнгать институтында, алар корпусны куллана. Калганнары, ялгышмасам, берән-сәрән эзләнүчеләр генә. Төрле илләрдән, төбәкләрдән эзләнүчеләр бар. Канададан, Европа илләреннән, Башкортстаннан, Себер якларыннан, Төркиядән кулланабыз дип язганнары бар. Киңәшләрен, үтенечләрен язалар.

- Көнгә сайтка ничә кеше керә?

- Анысын исәпләмибез. Чөнки безнең еллап кына исәпләү бара, көнләп исәпләүне эшләмәдек. Ул җәй көне кими бугай. Бу бит ялга кая барыйк микән дип эзләү системасы түгел, ә нәкъ менә эш башкару өчен система, шуңа күрә кеше күбрәк эш вакытында керә.

“Кулланучылар киңәшләрен җиткерсә иде”

- Гадәти кулланучылар корпуска ничек ярдәм итә ала?

- Без кулланучылардан ярдәм сорап сайтта да яздык инде. Әгәр аларда текстлар булса, бигрәк тә алар шул текстларның авторлары булсалар, безгә җибәрсеннәр иде. Корпусны кулланалар икән, анда электрон почтабыз күрсәтелгән, үз фикерләрен, тәкъдимнәрен язарга курыкмасыннар иде. Чөнки конференцияләрдә турыдан-туры менә бу мөмкинлегегез булса, әйбәт булыр иде, монысы сезнең ничегрәк эшли икән дип әйтәләр. Без корпуста кулланмалар да яздык, өч телдә, сайтыбыз да безнең өч телдә. Рус, татар, инглиз телләрен белүче кешеләр сайтның интерфейсын өч телдә дә күрә ала, ягъни аның белән эшли ала дигән сүз. Кулланмасы да өч телгә тәрҗемә ителде, анда 70-80 битлек кулланма. Ләкин шулай да сораулар кала, конференцияләрдә чыгыш ясаганда мөрәҗәгать итәләр. Очрашуларны көтмичә, тотыгыз да языгыз электрон почтага. Без башкалардан аерылып торган әллә нинди коммерцияле оешма түгел бит, кулланучылар белән бергә, алар өчен дип эшлибез. Алар үз фикерләрен әйтмәсәләр, корпусның файдасы да кимиячәк.

Билгеле, идеяләрне алардан гына көтеп калмыйбыз, идеяләр мәкаләләр укыганда да туа. Чөнки чит илләрдә һәм рус телендә бу өлкәдә бик зур эшләр башкарылды, без корпусны ясаганда алар тәҗрибәсенә таяндык. Мәсәлән, русларның бик әйбәт “Рус теленең милли корпусы бар”- “Национальный корпус русского языка”. Бик яхшы проект. Төрекләрең 500 миллион сүзне үз эченә алган зур корпуслары эшләнде. Аларның берничә корпусы бар һәм барысы да шактый әйбәт эшләнгән, сыйфатлы корпуслар, әле дә арта гына бара, мәкаләләрен укып барам. Алар Европа, Америка галимнәреннән артта калмаска тырышалар. Европа, Америка галимнәре чыгарган мәкаләләрне дә укыйбыз. Бигрәк тә Лейпциг университетында эшләнгән корпус турындагы материаллар файдалы булды, чөнки алар бик күп телләр өчен корпуслар ясыйлар. 230лап тел өчен 300гә якын корпуслары бар. Кечкенә генә татар теле корпусы да бар иде, бер-ике ел элек мин алар белән элемтәгә кереп, аларга татар теле корпусын зурайтырга ярдәм иттем.

“Татар телендәге корпуслар бик аз”

- Татар телендә ничә корпус бар?

- Бер кул бармаклары белән санап чыгып буладыр аларны. Sketch Engine дигән проект бар, ул, ялгышмасам, Чехия белән Англиядә эшләнә. Анда шулай ук бик күп телләр өчен күп корпуслар бар. Аларга да татар теле корпусын ясарга узган елны ярдәм иткән идем. Ләкин аларныкы түләүле система, Лейпцигныкы түләүсез.

- Татар телендәге корпуслар:
- Татар теленең язма корпусы: http://corpus.tatar/tt
- Россия фәннәр академиясе әзерләгән "Туган тел" корпусы: http://web-corpora.net/TatarCorpus/search/?interface_language=ru
- Leipzig Corpora Collection (Германия): http://corpora.uni-leipzig.de/en?corpusId=tat_mixed_2015
- Sketch Engine (Чехия-Англия): https://www.sketchengine.co.uk (түләүле)

Аны бит дөнья галимнәре кереп куллана ала. Мәсәлән, шул ук инглизләргә татар теле буенча ниндидер материал кирәк икән, һәм алар бу корпус белән эшли икән, шул системада эшләргә өйрәнеп беткән икән, ул безгә мөрәҗәгать итеп тормасын, безнеке аңа, бәлки, җайлы тоелмас. Шуңа күрә, андый галимнәргә җайлы булсын дип, мин аларга булышырга тырыштым.

Беренче карашка алар бит безнең конкурентлар, без аларга эшләштек икән, безгә кулланучылар азрак керәчәк дигән сүз. Ләкин, икенче яктан караганда, без үзебез акча эшләмәгәч, кулланучылар санын арттырырга дигән критерий юк. Без татар телен мөмкин кадәр күбрәк кулланучыларга, өйрәнүчеләргә җиткерик дигән ният белән эшлибез инде. Шуңа күрә монда бернинди дә эчкерлек юк, бер-беребезгә ярдәм итә алабыз икән, без һәрвакыт ачык.

“Корпуска керешкәндә барысы да кызыгып китәрләр дип уйлаган идем”

- Глобаль максатка килгәндә, татар теле корпусы нинди максатны күз алдында тотып эшләнде?

- Минем корпуска керешкәндә бер хыялым бар иде - эшли башларбыз да, безнең фәнни оешмалар, университетлар, фәннәр академиясе кызыгып китәрләр дип... Һәм, яңгырдан соң чыккан гөмбәләр кебек, компьютер лингвистикасы өлкәсендә башка проектлар да чыга башлар кебек тоела иде миңа. Кызганыч ки, бик авыр бирелә икән. Андый проектлар, берән-сәрән генә, авыр гына барлыкка киләләр, кайберләре шулай туктап та кала. Проект үсмәсә, туктап калдымы, бетте инде ул, әһәмияте шундук бетә дигән сүз.

Филология һәм мәдәниятара багланышлар институты директоры Рәдиф абый Җамалетдинов безнең бу проект белән кызыксынган иде, ул безнең белән очрашты, университетта да мондый проектлар башларга уйлап торалар, бәлки, башлаганнардыр да. Тел, әдәбият һәм сәнгать институты директоры Ким Мөгаллимович аның урынбасары Олег Ришатович белән бик кызыксынып, проектларында катнашырга чакырдылар. Эшне матур гына башлап җибәрдек, Алла бирсә, бәлки, ниндидер уңышларга ирешербез.

“Башкортлар бу өлкәдә алга киткән”

- Россиядә яшәүче бүтән кайсы халыкларның мондый корпуслары бар?

- Берничә ел элек Россия фәннәр академиясе илдәге бөтен телләр өчен корпуслар булдыру программасын эшләп чыгарды, анда зур акчалар билгеләнде, махсус сайт эшләнде. Шул проект нигезендә күп телләр өчен ниндидер корпуслар эшләнде. Аларның сыйфатын әйтә алмыйм. Ә менә аерым-аерым, РАН белән бәйле булмаган, ә үз җирлекләрендә эшләнгән проектлардан миңа комиларныкы бик ошады. Икенчедән, калмык теле корпусы матур гына эшләнеп килә иде, соңгы вакытларда караганым юк. Соңгы елларда башкортларның поэзия корпусы эшләнде. Башкортларда бу корпусларны эшләү күп еллар элек башланган иде һәм алар бу өлкәдә бик зур эш башкарганнар. Ялгышмасам, 1990 еллар азагындамы, 2000 еллар башындамы эшли башлаганнар. Аларның шул вакыттан бирле калган корпусларына юлыкканым бар. Ул, билгеле инде, бүгенге көн стандартлары буенча артта калганрак дип исәпләнергә мөмкин, ләкин шул көннәр өчен инде ул бик яхшы корпус булган. Бүгенге көндә дә аны бер горурланырлык әйбер дип уйлыйм.

- Корпусларның телне саклап калуда рольләре нинди?

- Тел өлкәсендә башкарылган теләсә нинди проектның күпмедер дәрәҗәдә телне саклауда роле

бардыр инде. Ярдәме күпме икәнен әйтергә авыр.

“Татар телендә профессиональ текстларны табу бик кыен”

- Профессиональ текстларны табу кыен түгелме? Мәсәлән, медицина өлкәсе буенча?

- Бик кыен. Әле күптән түгел генә конференциядә булган идем. Казан дәүләт архитектура һәм төзелеш университетында Шакирҗанов Рәшит абый һәм башкалар зур конференция оештырдылар, анда татар теленең техник фән өлкәсендә кулланылышы турында шактый гына кызыклы докладлар булды. Иң кызыгы - алар үзләре эшендә кирәкле әсбапларны татар телендә күп еллар дәвамында чыгарып киләләр. Бөтен терминнарны тәрҗемә итеп, дәреслекләр эшләнеп килә, бик күп башкарылган эшләре. Үзләре әйтүенчә, алар, нигездә, грантлар, фондлар финанслап түгел, ә институт кысаларында мөмкинлекләр табып эшләнә икән.

Шул конференциядә Рәшит абый Шакирҗанов уку йортында әзерләнгән зур техник терминнар сүзлеген (коллектив хезмәт) дә презентацияләде. Эшләргә теләүчеләр һәм эшләүчеләр ул өлкәдә дә бар икән. Татар теле фән өлкәсендә кулланыла икән һәм алар бу өлкәдә студентларны чынлап торып укыталар һәм мондый кадрлар әзерлиләр. Ел саен аларның күпмедер студент чыга. Соңгы 10-15 елда алар 400ләп татар телендә укытылган белгеч әзерләп чыгарганнар. Бу әсбапларның бөтенесен корпуска кертү мөмкинлеге булса, бу нинди байлык, уйлап карасаң.

“Бернинди грантлар да ота алмадык”

- Сезнең грантлар алырга омтылып караганыгыз булдымы?

- Әйтүемчә, эшнең башында акча турында уйламаган идек. Корпусны эшләп карыйк, булса - булыр, булмаса - юк дип башладык беренче версиясен һәм интернетка урнаштырдык. Шуннан соң безгә башкаларның да әйткәне булды, инде күрсәтер эшегез бар икән, грантка бирү авыр эш түгел, бирегез, корпусыгызны үстерегез диделәр. Төрле грантларга биреп карадык. Фәнни җитәкчем Тәүзих абый мәгариф һәм фән министрлыгына да ничә тапкыр мөрәҗәгать итте. Алар төрле конкурслар да оештырган иде бит, мәсәлән, ялгышмасам, “Татар телендәге иң яхшы интернет проект”, тагын ниндидер... Кызганыч ки, берсендә дә җиңә алмадык. Шуннан соң бер-ике ел тагын берничә грантка биреп карадык, 5-6 булды микән, артыграк иде микән. Берсендә дә үтә алмагач, бүтән бирмибез дигән уйга килдем. Чөнки грантка бирү өчен билгеле бер программалар төзисең, планнар язасың. Безнең корпус болай да буш вакытыбызда гына эшләнә. Үзеңне юкка, планнар төзеп утыруга да багышласаң, корпусның үзен эшләргә вакыт та калмый. Шуннан соң без сайтка язып та куйдык - бу проект бернинди акчага да эшләнми, бернинди финанс ярдәм дә алмыйбыз дип. Үзебез дә шуңа килештек - бүтән бернинди грантка да бирмәячәкбез. Узган ел безгә, кем икәнен әйтмим, мөрәҗәгать итүче дә булды, сезгә финанс ярдәм күрсәтәбез диделәр. Без баш тарттык, чөнки бүтән акчага эшләмибез дигән карар кабул иткән идек инде.

Безгә Telegramда язылыгыз

Әгәр хата күрсәгез, текстны билгеләгез һәм **Ctrl + Enter** төймәләренә басыгыз