Fledu.uz
foreign language education

Scientific-methodological electronic journal
"Foreign Languages in Uzbekistan", 2020, No 1(30), 50-58
https://journal.fledu.uz
ISSN: 2181-8215 (online)

ЗАМОНАВИЙ ЛИНГВИСТИК КОРПУСЛАРНИНГ КОМПЬЮТЕР МОДЕЛЛАРИ

Нилуфар Зайнобиддин қизи АБДУРАҲМОНОВА
Филология фанлари бўйича фалсафа доктори (PhD)
Алишер Навоий номидаги Тошкент давлат ўзбек тили ва адабиёти университети
Тошкент, Ўзбекистон

КОМПЬЮТЕРНЫЕ МОДЕЛИ СОВРЕМЕННОГО КОРПУСА

Нилуфар Зайнобиддин кизи АБДУРАХМОНОВА
Доктор философии (PhD) по филологическим наукам. Ташкентский государственный университет узбекского языка и литературы им. Алишера Навои. Ташкент, Узбекистан

COMPUTER MODELS OF CONTEMPORARY CORPORA

Nilufar Zaynobiddin kizi ABDURAKHMONOVA
Doctor of Philosophy (PhD) in Philological Sciences. Tashkent State Uzbek Language and Literature University named after Alisher Navoi. Tashkent, Uzbekistan
abdurahmonova.1987@mail.ru

UDC (УЎК, УДК): 81–139

For citation (иқтибос келтириш учун, для цитирования):
Абдураҳмонова Н.З. Замонавий корпусларнинг компьютер моделлари // Ўзбекистонда хорижий тиллар. — 2020. — № 1(30). — Б. 50–58.
https://doi.org/ 10.36078/1583734626

Received: December 27, 2019
Accepted: February 10, 2020
Published: February 20, 2020

Аннотация. Сўнгги вақтларда матн корпуслари дунё компьютер лингвистикаси (NLP) ва тилшуносликнинг турли соҳалари учун энг муҳим ўрганиш объектига айланиб бормоқда. Бироқ ўзбек тилшунослигида корпус билан боғлиқ жиддий амалий тадқиқотлар амалга оширилган эмас. Шу боис ушбу мақолада корпуснинг компьютер моделларини яратишга доир изланишларни таҳлил қилиш ва улардан унумли фойдаланиш учун эришилган натижаларни қиёсий ўрганишга эътибор қаратилган. Мақолада ўрганиш объектининг ўзига хос хусусиятлари ва фойдаланувчиларнинг турли мақсадларидан келиб чиқиб, электрон корпусларнинг шакллантириш босқичлари бир нечта корпусларнинг қиёсий таҳлили асосида ўрганилди. Таҳлил натижалари шуни кўрсатадики, матнга доир метамаълумотларнинг берилиши, лингвистик жиҳатдан аннотациялаш ва жанрлар таснифини мувофиқлаштириш барча корпуслар учун умумий жиҳатларидан биридир. Компьютер лингвистикасида корпус яратиш меъёрлари ва мезонларини аниқлаш кейинги тадқиқотлар учун муҳим лингвистик ресурс бўлиб хизмат қилади. Тадқиқотнинг натижалари асосида қуйидаги хулосага келинди: корпуснинг морфологик, синтактик ва семантик аннотациялари ёрдамида компьютер лингвистикасининг турли илмий йўналишларида дастурий таъминотлар (маълумотларни классификациялаш, маълумотларни қайта ишлаш, машина таржимаси, сентимент анализ) яратиш учун муҳим манба ҳисобланади.
Калит сўзлар: корпус; компьютер моделлари; корпус лингвистикаси; аннотациялаш; компьютер лингвистикаси.

Аннотация. В последнее время корпус текстов стал важнейшим объектом изучения для мировой компьютерной лингвистики (NLP) и различных областей лингвистики. Однако, до настоящего времени, практически не было исследований в области узбекской корпусной

лингвистики. Данная статья посвящена анализу компьютерного моделирования корпусных исследований и сравнительному изучению результатов, достигнутых при их использовании. В статье рассматриваются этапы построения электронных корпусов в соответствии с потребностями пользователей и специфическими особенностями объекта исследования путем сравнительного анализа. При этом текстовые метаданные, лингвистическая аннотация и жанровая классификация считаются общими характерными чертами для всех существующих корпусов. Определение принципов и критерии создания корпуса в компьютерной лингвистике служат важнейшем языковым ресурсом для дальнейших исследований. На основании результатов исследования, был сделан вывод о том, что морфологические, синтаксические и семантические аннотации корпуса являются важными источниками для разработки программного обеспечения в различных научных областях компьютерной лингвистики (классификация данных, обработка данных, машинного перевода и сентимент анализа).
Ключевые слова: корпус; компьютерные модели; корпусная лингвистика; аннотация; компьютерная лингвистика.

Abctract. Recently the corpus of texts has become the most important obje of study for worldwide computational linguistics (NLP) and the various fiel of linguistics. However there has not been practical implementation research in the area of Uzbek corpus linguistics. Therefore much attention given in the article to analysis of investigations in the scope work on buildi computational models of corpus and observing comparatively achiev results for efficient usage of them.
The paper deals with the stages of building electron corpora according to different purposes of users and specific peculiarities of an object of study by comparative analysis. In this case balancing genres classification and linguistic annotation, and representation metadata considered common characteristic features for all existing corpora.
Identifying principles and criteria of creating corpus in computational linguistics serve as crucial linguistic resource for the further researches. Based on the results of the study, it was concluded that morphological, syntactic and semantic annotations of the corpus are important sources for software development in various scientific areas of computational linguistics (data classification, data processing, machine translation, sentiment analysis and so on).
Keywords: corpus; computational models; corpus linguistics; annotation; computational linguistics.

I. Корпус лингвистикасига кириш. Тил ҳамиша кишилик жамиятнинг энг ажралмас алоқа воситаси бўлиб хизмат қилади. Унинг ёрдамида янги билимга эришилади ва эгалланган билим тафаккур ва тажриба асосида қайта ишланади. Бугун тил саноатида компьютер лингвистикаси, машина таржимаси, тил технологияси, табиий тилни қайта ишлаш, сунъий интеллект технологияси каби терминлар тез-тез қўлланилиб келинмоқда. Буларнинг ҳар бири инсоният томонидан яратилган лингвистик ресурс орқали ўз тадрижий такомилига эга. Эндиликда лингвистик ресурсларнинг асосий турларидан бири сифатида электрон шаклидаги корпуслар назарда тутилмоқда.

Корпус лингвистикаси матнларни йиғиш, таснифлаш, аннотациялаш каби вазифаларни бажаради. Корпус машина учун табиий тилни тушуниш жараёнини лингвистик жиҳатдан аниқ ва тўғри кўрсата олгани боис табиий тилни қайта ишлаш (NLP-natural language processing) соҳасида унинг ўрни муҳим аҳамиятга эга. Шу боис илмий

манбаларда корпус лингвистикаси табиий тилни қайта ишлаш соҳасининг (NLP) бир йўналиши сифатида қайд этилади (4, 12). Айрим маълумотларда эса компьютер лингвистикаси ёки амалий лингвистиканинг муайян соҳаси сифатида эътироф этилади.

NLP тизими учун лингвистик билимлар жуда муҳим. Улар қабул қилинган муайян лисоний моделлар, қоидалар, луғатлар шаклида ифода этилсада, бироқ тилнинг нутқий омили билан боғлиқ бўлган экстралингвистик маълумотларга ҳам эҳтиёжи катта. Шу маънода тилнинг концептуал инъиқоси акс этган семантик тармоқлар, онтология кўринишидаги билимлар таъминоти зарур. Демак, тилнинг мураккаб функционал имкониятларини прагматик, нейролингвистик, психолингвистик ёки дискурсив параметрларини баҳолашда NLP соҳасининг олдида жуда мураккаб босқичларни бажариш вазифаси мавжуд. Шунинг учун ўзбек тилшунослигининг жуда муҳим вазифалардан бири ўзбек тили миллий корпусини яратишдир.

Ушбу мақолада корпус лингвистикасида эришилган ва натижаси кутилаётган амалий ва назарий ёндашувлар асосида корпус моделлари бўйича муайян таҳлилий муносабатлар тадқиқ этилган.

Калифорния университети мутахассиси Стефан Гриес ўзининг илмий қарашларида корпус учун берилган таърифлар “У метод, назария ёки моделми?”, — деган саволга нисбатан уни *метод(ология*) деб баҳолайди. Фикрини асослаб, тил назариясининг генератив нуқтаи назаридан тилга алоқаси йўқ, деган тўхтамга келади. У дескриптив ва амалий жиҳатдан айрим методлардан фойдаланиб мисоллар ёрдамида ўз фикрини исботлашга ҳаракат қилади. Унга кўра агар тилшунос бирор лексемани корпус ичида грамматик бирлик сифатида ўрганмоқчи бўлса, демак у грамматик назарияга, агар у ёки бу миллатнинг иккинчи тил сифатида мураккаб конструциялардан қай даражада қўллаш имкониятини баҳоламоқчи бўлса, иккинчи тилни ўрганиш назариясига асосланади, яъни ҳар бир изланувчи ўзига хос методни танлайди, деган муносабатни ўртага ташлайди.

Чарлотте Тайлор “Корпус лингвистикаси нима?” сарлавҳали мақоласида корпус лингвистикаси ҳамиша частотага асосланиши, ундаги маълумотлар веб саҳифаларидаги матнлардан ташқари нутқий жараёнда яратилган турли услубдаги ёзма ва оғзаки материаллар (газета ва журнал материллари) ҳамда аудио кўринишидаги маълумотлардан ташкил топиши билан фарқланишини қайд этади. Шу жиҳатдан тадқиқотчи корпуснинг умумий ва махсус турлари мавжудлиги, махсус турдаги корпуслар жанр, услуб, даврларга кўра фарқланиши, ҳар икки турдаги корпус ўз навбатида диахрон ва синхрон шаклида бўлиши мумкинлигини қайд этади. Олим ўзининг илмий қарашларида диахрон корпус тилнинг даврлар оша қай даражада ўзгарганлигини глоттохронологик ва статистик таҳлил қилиш учун асосий объект бўла олиши, синхрон корпус эса айни вақтда қўлланилаётган нутқий бирликларнинг замонавий қўлланилишини акс этишга ёрдам беришини айтиб ўтади. Шунингдек, Ч. Тайлор изланишларида корпус монолингвал ва параллел матнларнинг мажмуасидан иборат бўлиб, турли соҳалар учун махсус ўрганиш соҳаси бўла олишини кўплаб мисолларда исботлашга ҳаракат қилди.

Чарлез Меер корпус лингвистикасида тадқиқот олиб бораётган муайян кичик доирадаги тадқиқотга доир методологик усулни ҳам корпус деб номлаш мумкинлиги таклиф этади. Олим ўз қарашларида мақолаларнинг онлайн корпуси (Маниез, 2000) яратилганлиги, бироқ элекрон шаклидаги мақолалар корпусми ёки матнлар корпусидан олинган мақолалар унинг таҳлил натижасими, деган саволни ўртага ташлайди. У корпусга таъриф беришда тил инженерияси соҳасининг стандартлар бўйича экспертларининг маслаҳат гуруҳи (The Expert Advisory Group on Language Engineering Standards — EAGLES) умумий тарзда корпус сифатида нафақат насрий газета, назмий драма турдаги матн турлари, балки сўз рўйҳати ва луғатлар ҳам мисол бўла олишини эътироф этади (6). Ушбу таърифга кўра, барча лингвистик манбаларни корпус сифатида қараш мумкин, дейди Ч. Меер (3, 12). Олим корпус матнларнинг жамланмасидир, деган фикрга нисбатан ушбу мулоҳазани билдиради: Отто Жесперснинг кўп серияли тарихий тамойиллар асосидаги “Замонавий инглиз тили грамматикаси”да Чаусер, Шекспир, Свифт, Аустин ва Жесперсенларнинг асарларидаги матнлардан турли лингвистик структуралар қиёсланган. Бироқ бу корпусга асосланган таҳлил дейилмасада, лекин у ўзининг тадқиқотида айнан корпус вазифасига ўхшаш матн фрагментидан фойдаланилган. Демак, электрон корпуслардан олдин, матнлар жамланмасидан турли лингвистик мақсадларда фойдаланилган.

Маълумки, нутқнинг ёзма ва оғзаки шакллари бўлиб, унинг индивидуал тарзда ранг-баранг ифодаланиши турли услубларда ўз аксини топади. Тилнинг бу имконияти унинг чексизлигини кўрсатади. Дунёда нечта тил бўлса, у миллионлаб миллат ва элат нутқида прагматик кўринишига эга. Сўзловчиларнинг турли омилларга кўра фарқланиши унинг ўзига хос жиҳатларидан биридир. Корпус лингвистикасида нутқнинг барча ифодасини муайян даражада кузатиш, таҳлил қилиш, ўрганиш имконияти бор. Корпус матнлар мажмуи сифатида ўрганилаётган объект ва предметнинг тизимли мажмуасидир. Бу борада В. Захаров: “корпус тилнинг қисқартирилган модели”, деб ўринли баҳо беради. Зеро тилнинг турли дискурсдаги ифодаси табиий нутқ ҳолатида намоён бўлади.

Шу маънода корпуснинг турли компьютер моделлари яратилган бўлиб, унинг яхлитлиги қўйилган мақсад ва вазифалардан келиб чиқади, келишилган тамойиллар ва аудитория шароитига мослаштирилади. Шунинг учун корпус концепциясини тўғри англаш муҳим масалалардан биридир. Ноам Чомскийнинг корпусга доир билдирган қарашларида айни ҳақиқат мавжуд. Унга кўра тилнинг чекли қоидалари бўлишига қарамай, чексиз жумлалар яратиш имконияти бор. Тил жамиятда турли социал омиллар таъсирида, неологизмларнинг кириб келиши натижасида корпус нечоғлик катта ҳажмга эга бўлмасин тилда акс этувчи барча нутқий имкониятларни бир вақтнинг ўзида жамлай олмайди. Бизнингча, олимнинг матн корпусига билдирган фикрлари айни ҳақиқат. Негаки корпус ҳар қанча катта бўлмасин тилнинг яхлит манзарасини тасвирлашга қодир эмас.

Таъкидланганидек, корпус ёки корпус лингвистикасига олимларнинг нуқтаи назарлари турлича ёндашилган. Қайд этилишича, у тил модели эмас, уни муайян даражада методологик ёндашув сифатида қараш

тўғри бўлади. Бу борада Доуглес Бибер корпуснинг қуйидаги хусусиятларини санаб ўтади (2, 193):

— у табиий матндаги зарур бирликларни эмпирик таҳлил қилади;

— анализ учун *корпус* сифатида табиий матнларнинг катта ва тизимлаштирилган жамланмаларини бирлаштиради;

— анализ учун компьютернинг ҳам автоматик ва интерактив технологияларидан фойдаланиш имконини беради;

— у аналитик технологиянинг миқдор (статистик) ва сифат хусусиятларини ўз ичига олади.

Тил даврий силсиланинг маҳсули сифатида жамият лабораториясининг маҳсули бўлиб қолаверади. Чунки тилимиздаги услубий ўзига хослик муайян йўналишда ўз позицион муҳитида янгилик қилишдан тўхтамайди. Натижада корпусни узлуксиз равишда бойитиб, янгилаб бориш талаб қилинади. Бу эса корпусшунослик соҳасида лингвистик тадқиқотлар изчиллик билан давом этиш зарурлигини кўрсатади.

Эндиликда корпуснинг энг мақбул шакли рақамли (электрон) кўриниш бўлиб, машина (компьютер)да қайта ишлаш, таҳлил қилишнинг осон, қулай ва тезкор усули сифатида корпуснг универсал, стандарт форматда бериш ҳисобланади. 1990 йилга келиб дунё тилларининг компьютер анализига мўлжалланган 600га яқин корпуси борлиги аниқланган (5, 12).

Бугунги кунда корпуснинг имкониятлари шу даражага етдики, ушбу соҳада эришилган ютуқлардан нафақат NLP соҳаси ёки компьютер лингвистикаси, балки тилга ўқитишнинг методика ва педагогика соҳалари, дискурс таҳлил, машина таржимаси ва лингвистиканинг ўнлаб соҳалари (социолинггвистика, гендершунослик, психолингвистика ва ҳ.к.) унумли фойдаланиб, ижобий натижага эришиб келмоқда.

I. **Корпус таксонимияси.** Белгиланган мақсадга кўра корпуснинг турли моделлари мавжуд ҳамда уларни яратишда турли меъёрларларга амал қилинади:

1. Берилганлар базасининг турларига кўра: оғзаки, ёзма, аралаш;
2. Муайян тилда матннинг берилиши: инглиз, рус, немис;
3. Матн таржималарининг параллелигига кўра: икки тиллик, уч тиллик, кўп тиллик;
4. Услубига кўра: сўзлашув, публицистик, бадиий, расмий, илмий;
5. Базадан фойдаланиш имкониятига кўра: очиқ, ёпиқ;
6. География ҳолатига кўра: фақат бир давлатга тегишли ва ҳ.к.

Умумий фойдаланиш мақсадидан келиб чиқиб корпус икки турга ажратилади:

1. Умумий корпус — бу каби корпусда матннинг турли жанрлари ва ифода шакллари (оғзаки ва ёзма) аралаш ҳолда кенг қамровли бўлади;

БЕЛГИСИ	КОРПУС ТУРИ
Мақсади	кўп мақсадли; муайян мақсадга йўналтирилган
Тилга оид маълумотларнинг тури	оғзаки; ёзма; аралаш
Адабий тилга оидлиги	диалектал; сўзлашув; терминологик; аралаш
Жанр	бадиий; фольклор; драматург; публицистик
Кўриниши	тадқиқотга оид; иллюстратив
Ўзгарувчанлик	динамик; барқарор
Лингвистик аннотация (разметка)	разметкаланган; разметкаланмаган
Аннотация (разметка) тури	морфологик; синтактик; семантик; анафорик; просадик
Фойдаланиш имконияти	барча учун очиқ; ёпиқ; тижорат мақсадли
Матн ҳажми	тўлиқ матнли; матндан берилган парча (матн фрагменти)

жиҳатдан анчагина кичик матнлар мажмуидан иборат бўлади.

В. Захаров корпусни қуйидагича таснифлайди: “Қайд этилган оғзаки нутққа хос бўлган матнлар корпуси хусусиятига кўра сўзларнинг талаффузи (транскрипцияси) ва турли диалогларда қўлланилган нутқий вазиятлар келтирилади. Бу каби корпуслар таълим жараёнида, яъни у ёки бу тилни чет тили сифатида ўқитишда муҳим аҳамият касб этади” (5, 16). Корпуснинг ёзма шакли турли манбалардан олинади: веб саҳифалар, ёзма материаллар, газета ва журналлар, диний манбалар ва ҳ.к.

Хуллас, ёзма матн кўринишидаги барча электрон ва қоғоз ҳолдаги ресурслардан корпус сифатида фойдаланиш мумкин. Оғзаки матн корпуси М. З. Курди нуқтаи назарига кўра қуйидагича бўлади (4, 4):

— **Оғзаки ёзиб олиш.** Бунда махсус йиғилган матнлар турли категорияга кирувчи (ёш, жинс, касб, нутқий имконият, шева ва ҳ.к.) кишилар томонидан ўқитиб олиниб, улардаги фонетик ўзига хослик ёзиб олинади;

— **Гапирувчи буйруқ операторлари.** Бунда оғзаки матнлар телевизор ёки робот орқали ёзиб олинади. Ушбу ҳолатда табиий муҳит ўзгаради, чунки унда матннни беихтиёрий бузиш, сўзларни тез-тез такрорлаш ёки тушуриб қолдириш каби ҳолатлар кузатилмайди.

— **Инсон-машина диологи.** Бунда машинанинг имконияти анча чекланган бўлиб, инсон машинага мослашиш учун содда бирликлардан фойдалангани боис, бунда лингвистик феноменнинг хилма-хилиги йўқолади;

— **Машина ёрдамидаги инсон — инсон диологи.** Ушбу ҳолатда оғзаки ёки ёзма матн икки киши томонидан яратилсада, машина ундаги ёзишмаларни кетма-кетликда ёзиб олади. Айниқса, таржима тизимида бу жараёндан фойдаланилади.

Кўп моделли диолог. Диолог инсон билан инсон ёки инсон билан машина ўртасида бўлса-да машина улар ўртасидаги мулоқотни таъминлашга воситачи бўлади. Бундай диологларда сўзлар ва ишоралар аралаш тарзда ишлатилади.

Корпусларнинг тилига кўра монолингвал, билингвал ва кўп тилли бўлиши тадқиқотчи томонидан ўрганилаётган объектнинг хусусиятларидан келиб чиқади. Дунё бўйича жуда кўп тилларнинг корпуслари яратилган: Brown корпуси, инглиз тилининг ёзма ва оғзаки матнлар корпуси Switchboard корпуси, француз тили учун Frantext корпус, французча оғзаки

ва ёзма матнлар учун OTG корпусини алоҳида қайд этиш ўринли бўлади. Параллел корпус бир неча тилдаги матнлар мажмуасидан иборат бўлиб, улар маълумотлар базасида ўзаро боғланади. Масалан, расмий сайтлардаги маълумот ёки янгиликлар икки ёки ундан ортиқ тилларда берилади. Бу ўз навбатида матннинг таржима муқобиллари тарзида маълумотлар базаси (МБ) сифатида йиғилиб машина таржимаси учун ресурс бўлиб хизмат қилади. Мултилингвал (кўп тилли) корпусларнинг тўплами параллел корпуслардан иборат бўлмай, берилган матн бошқа тилда айнан ўхшаш бўлмайди. Намуна сифатида учта диалект ва ўн икки тилдаги телефон суҳбатларидан ташкил топган CALLFRIEND корпуси ва олти тилдаги диалоглардан таркиб топган CALLHOME корпусини келтириш мумкин (4, 7).

М. З. Курди корпусларнинг тематик кўламини матнларнинг соҳавий таснифига кўра уч турга ажратади:

- мувозанатлашган корпус — бунда олинган тематик мавзу тенг миқдорда тақсимланади;
- пирамидали корпус — бунда мавзунинг долзарблигига кўра марказий ва кичик матнлар мажмуасидан иборат бўлади;
- имконият даражасидаги корпус — лингвистик ресурси етарли бўлмаган ҳолатда берилган тил ёки илова учун яратилган бўлади.

Исталган турдаги корпус яратилишида қуйидаги саволларга жавоб изланиши керак: 1. Корпус қандай қилиб яратилади? 2. Қайси турдаги матнлардан тузилади? 3. Нима мақсадда фойдаланилади? 4. Имкониятлари нималардан иборат бўлади?

Сўнг йиғилган матнлар муайян ҳолатда аннотацияланиб, муайян мақсадга йўналтирилади. Демак, дастлаб корпус яратиш учун матннинг шакли (оғзаки ёки ёзма), ҳажми, услуби ва тематикаси аниқлаб олиниши керак.

Инглиз тили учун корпуслар платформаси яратилган бўлиб (6), юқорида тилга олинган корпуснинг умумий концепциясини мавжуд электрон корпуслардаги имкониятлари мисолида қиёсий таҳлил қиламиз.

Америка миллий корпуси (7) (Open American National Corpus (OANC)) 1990 йилдан буён амалда қўлланилиб келинмоқда. Унинг ҳажми 15 млн. сўзни ташкил қилади. Корпус қуйидаги бирликлардан ташкил топган: *DOCUMENT NAME* асосий маълумот файли учун берилган ном ва барча аннотацияларнинг боғланган маълумотлари. *WORD COUNT* ҳужжатлардаги мавжуд сўзларнинг миқдори, *TYPE* матннинг ёзма ёки оғзаки шакли, *GENRE* матнни таснифлаш учун корпуснинг луғатидан олинган маълумотларнинг жанри, *SUB-CATEGORY* мавжуд маълумотларнинг қўшимча таснифи, *ANNOTATIONS* мантиқий тузилиши (матннинг асосий қисми, параграфи, сарлавҳаси)дан асосий тўпламдан иборат ҳужжатлар; жумлаларнинг чегаралари, токен ва сўз туркумлари, от ва феъл сўз туркумини ҳамда номлар (шахс, ҳудуд, сана ва ташкилот)ни аниқловчи аннотациялар мавжуд. Ушбу корпусда 19 турдаги жанрга оид маълумотлардан фойдаланилган бўлиб, корпусдаги мавжуд имкониятлар муайян лойиҳалар доирасида олиб борилган тадқиқотлар натижаси саналади.

Оксфорд инглиз тили корпуси (8) (Oxford English Corpus (OEC)) Оксфорд университети нашриёти учун лойиҳалар устида иш олиб бораётган ходимлар учунгина фойдаланиш имконияти мавжуд. Ундаги маълумотлар 2000–2006 йилдаги ҳужжатларни қамраб олсада ҳажм жиҳатдан улкан корпус, деб баҳоланади (9). Унда 2.1 млрд. сўз бўлиб, Буюк Британия, АҚШ, Ирландия, Австралия, Янги Зинландия, Кариб ороллари, Ҳиндистон, Канада, Сингапур ва Жанубий Африкадаги истеъмолдаги инглиз тилидаги ҳужжатлардан ташкил топган. Уларнинг асосий қисми веб-саҳифалардан,

айримлари босма нашрлар ва турли соҳага алоқадор илмий журналлардан иборат. Унда берилган файллар XML форматда бўлиб, Sketch Engine дастурий таъминоти ёрдамида анализ қилинади. Ҳар бир ҳужжатнинг метамаълумоти қуйидаги мундарижадан иборат: сарлавҳа, муаллиф, муаллифнинг жинси, тили (Британия ёки Америка инглиз тилиси), веб саҳифаси манбаси, йили, соҳаси ва ички соҳалари, ҳужжат статистикаси (токен, жумлаларнинг миқдори ва ҳ.к.).

Замонавий Америка инглиз тили корпуси (The Corpus of Contemporary American English (COCA)) 1990–2017 йилдаги мавжуд турли жанрдаги 560 млн. сўзли жанр жиҳатдан тенг миқдорда мувозанатлашган оғзаки матнлар, бадиий адабиёт, оммобоп журналлар, газета ва илмий матнлардан ташкил топган.

Рус тилининг миллий корпуси (12). 2013 йилги маълумотга кўра, ушбу корпус 500 млн.га яқин сўзни ташкил этган. Матнлар илмий, расмий, оммабоп, бадиий ҳамда сўзлашув услуби, электрон мулоқот матнлари ҳамда мультимедия маълумотларидан иборат. Мультимедияли рус корпуси 1930–2000 йилларда яратилган кинофильмлар фрагментидан тузилган.

Хелсинки аннотацияли корпус (11) (ХАНКО-Хельсинкский Аннотированный Корпус). 2000 йил бошида Хелсинки университетида ушбу корпус “Рус тилининг функционал синтаксиси” лойиҳасининг бир қисми сифатида яратилган. Мунтазам ривожлантирилиб бориладиган ушбу корпусда «Итоги» журналида 2001–2012 йилгача эълон қилинган мақолалар асосидаги ҳужжатлар 100 минг сўзшаклни ўз ичига олади. Корпусда морфологик ва синтактик қидирув тизими мавжуд. Корпуснинг ўзига хослиги маълумотларнинг лингвистик тавсифи изчил ишланган формат ҳамда автоматик морфологик ва синтактик разметкалар натижаларини тўлиқ визуал текшириш имконияти мавжудлигидир. Унда 2000та кўп таркибли барқарор бирикмалар киритилган.

Рус тилининг миллий корпуси (10) (Национáльный кóрпус рýсского языкá (НКРЯ)) 2004 йилда онлайн тарзда фойдаланила бошланган. Ушбу корпус XVIII аср ўрталаридан XXI аср бошигача бўлган даврий жиҳатдан таснифланган манбаларни қамраб олади. Ушбу корпус ўтган ва ҳозирга давр ўртасидаги социолингвистик жиҳатдан тартибланган адабий, диалектал, сўзлашув тилига хос бўлган турли адабий (драма, проза, поэзия) турга оид адабиётлар, шунингдек, очерк, эссе, журналистик нашр, илмий ва оммабоп илмий адабиёт, коммуникатив нутқ, ёзишма ва кундалик ҳужжатлардан тузилган 283 млн сўзни ташкил этади. Эндиликда ушбу платформа Яндекс компанияси томонидан қўллаб-қувватланиб келинмоқда.

Хулоса. Умуман олганда, корпус тилдан фойдаланишда унинг статистик анализи, табиий тилни қайта ишлаш (NLP) дастурий таъминоти, лексик ресурларни яратиш, тил ўқитишда ёки ўрганиш каби мақсаларда қўлланилади. Матнлар корпуси тилнинг динамик ҳолатини тадқиқ қилишда ёки лингвистиканинг турли соҳа предметига кўра анализ қилишда муҳим объект ҳисобланади. Корпус лингвистикасининг вазифаларидан бири лингвистик ресурсларнинг компьютерли анализи ва базасини яратишдан иборат. Шу боис унинг ёрдамида компьютер лингвистикасининг турли илмий йўналишларида дастурий таъминотлар (маълумотларни классификациялаш, маълумотларни қайта ишлаш, машина таржимаси, сентимент анализ) яратиш учун қўлланилади, шунингдек, корпуснинг морфологик, синтактик ва семантик аннотациялар ёрдамида лингвистик анализ учун муҳим электрон ресурс бўлиб хизмат қилади.

Фойдаланилган адабиётлар
1.Anke Lüdeling, Merja Kytö Corpus Linguistics An International Handbook,Vol. 1, Berlin, New York: Walter de Gruyter. 2008. 81(2), — P. 246–247 DOI: 10.1080/00393270903392342

2.Bern Heine, Heiko Narrog. The Oxford Handbook of Linguistic Analysis. / Dougles Biber Corpus-based and Corpus-driven analysis of language variation and use UK: Oxford university, 2015 — 193 p.
3.Charlez Meyer English corpus linguistics: An introduction. Cambridge University Press, 2004. — 168 p.
4.Mohamed Zakaria Kurdi. Natural Language Processing and Computational Linguistics: Speech, Morphology and Syntax. — Great Britain, USA: Wiley-ISTE 2016. — 300 p.
5.Захаров В. П., Богданова С. Ю. Корпусная лингвистика/Учебник для студентов гуманитарных вузов. — Иркутск: ИГЛУ, 2011. — 161 с.
6. https://www.english-corpora.org/
7. http://http//www.anc.org/data/masc/corpus/
8. https://corpus.byu.edu/coca/old/help/compare_oec.asp
9. https://en.wikipedia.org/wiki/Oxford_English_Corpus
10.https://www.google.com/search?q=http%2F%2Fruscorpora.ru%2Fsearch-
11.http:// www.ling.helsinki.fi/projects/hanco/
12.http://ruscorpora.ru

References
1.Anke Lüdeling, Merja Kytö *Corpus Linguistics an International Handbook*, Vol. 1, Berlin, New York: Walter de Gruyter. 2008. 81(2), p. 246–247 DOI: 10.1080/00393270903392342
2.Bern Heine, Heiko Narrog, *The Oxford Handbook of Linguistic Analysis*, Dougles Biber Corpus-based and Corpus-driven Analysis of Language variation and use UK: Oxford university, 2015, 193 p.
3.Charlez Meyer *English corpus linguistics: An introduction.* Cambridge University Press, 2004. 168 p.
4.Mohamed Zakaria Kurdi, *Natural Language Processing and Computational Linguistics: Speech, Morphology and Syntax*, Great Britain, USA: Wiley-ISTE, 2016, 300 p.
5.Zakharov V. P., Bogdanova S. Yu. *Korpusnaya lingvistika*, Irkutsk: IGLU, 2011, 161 p.
6. https://www.english-corpora.org/
7. http://http//www.anc.org/data/masc/corpus/
8. https://corpus.byu.edu/coca/old/help/compare_oec.asp
9. https://en.wikipedia.org/wiki/Oxford_English_Corpus
10. https://www.google.com/search?q=http%2F%2Fruscorpora.ru%2Fsearch-
11. http:// www.ling.helsinki.fi/projects/hanco/
12. http://ruscorpora.ru