Қазақ тілінің ұлттық корпусындағы негізгі ұғымдар

1. Терминдер мен қысқартулар

ҚТҰК ПМК – Қазақ тілінің ұлттық корпусының публицистикалық мәтіндер кіші корпусы

Корпустық лингвистика – лингвистикалық корпустардың компьютерлік технологиялар арқылы құрылуы мен қолданылу мәселелерімен айналысатын ғылым.

Корпус – машиналық оқылым форматындағы мәтіннің туу жағдайы, айтушы, автор, адресат немесе аудитория жөніндегі ақпаратты қамтитын мәтіндер жиынтығы.

Жалпы корпус – жалпы тілдің қолданысын көрсететін корпус.

Арнайы корпус – тілдің тар (не жеке) қолданыс аясын бейнелейтін корпус.

Параллель корпус – мәтіндердің бір тілдегі және екінші тілдегі аудармасымен қатар берілетін корпусы.

Синхронды корпус – тілдің белгілі уақыт шеңберіндегі күйін бейнелейтін мәтіндер корпусы (немесе белгілі мерзімге қатысты).

Диахронды корпус – тілдің әртүрлі даму кезеңдерін бейнелейтін мәтіндер корпусы.

Метадеректер – мәтін мен оның құрамдас бөліктері туралы қосымша.

Конкорданс - іздеген сөз кездесетін сөйлемдер легін шығару функциясы

Леммалау (лемматизация) – конкорданстағы кез келген сөзформаны түбір мен қосымшаға автоматты түрде бөлу;

Лингвистикалық белгіленім – тіл деңгейлеріне сәйкес білгіленім жасау.

морфологиялық белгіленім – сөздің морфологиялық құрамын сипаттау;

сөзжасамдық белгіленім – сөз жасаушы тұлғаны сипаттау;

лексикалық белгіленім – сөз мағыналарын белгілеу;

фонетикалық белгіленім - дыбыстар сипаттамасы және автоматты буынға бөлу;

морфо-семантикалық белгіленім - сөзге қатысты семантикалық категориялар мен оның мағынасына тән ішкі категорияларды сипаттау.

2. Корпус базасына енгізілген мәтіндердің белгіленімділігі

Линвистикалық белгіленім түрлері

Лингвистикалық белгіленім – сөздер мен сөзформалары элементтерін фонетикалық, морфологиялық, семантикалық, синтаксистік сипатын белгілейтін параметрлер жиынтығы. Жартылай автоматтандырылған, жартылай механикалық түрде енгізу арқылы жүзеге асырылған.

Қазақ тілінің ұлттық корпусының мәліметтер қорындағы сөздер мен сөзформаларды сипаттау үшін төмендегі белгіленімдер қолданылды:

- **Фонетикалық** – сөздердің дыбыстық құрамын, дыбыстардың сипатын, буын жігін автоматты түрде айырып, бөледі. Фонетикалық белгіленім іздеген сөзді мәтіннен көрсеткен кезде ең жоғары бөліктен көрініп тұрады.
- **Семантикалық** белгіленім – белгілі бір сөз табына жататын сөздің сол сөз табы бойынша морфо-семантикалық категориялық сипатын көрсетіп тұрады.
- **Лексикалық белгіленім –** сөз мағынасын, қажет жағдайда бірнеше мағыналарын беруді көздейді. Омоним сөздердің барлық мағыналары лексика бөлігінде көрсетіледі.
- **Морфологиялық**: Көптеген ірі корпустар морфологиялық белгіленімді болып келеді. Морфологиялық белгіленім әрі қарай синтаксистік және семантикалық талдау түрлерін жүзеге асыру үшін негіз болып табылады. Морфологиялық белгіленім сөз таптары мен сол сөз таптарына қатысты категорияларды айқындап, белгілеуден тұрады. Морфологиялық белгіленімнің автоматты бағдарламалық құралы – тэггер (taggers). Морфологиялық белгіленімді автоматтандыру мақсатында морфологиялық көрсеткіштердің базасы жасалады.

Сөз формасының морфологиялық құрылымы – лексеманың немесе лемманың атауы, сөз табына қатысы, морфологиялық сипаттамалары, яғни тиісті морфологиялық категорияларын анықтау жатады.

Синтаксистік белгіленім – лексикалық бірліктер мен синтаксистік конструкциялар арасындағы синтаксистік байланыстарды сипаттайды. Синтаксистік талдаудың автоматты бағдарламалық құралы – парсер (parsers).

Ескерту: Корпус базасындағы мәтіндерге синтаксистік белгіленім жасалған жоқ.

Корпусқа салынған мәтіндердің метамәтіндік белгіленімдері қажетті сөзді іздеу кезінде меңзерді мәтіннің жоғарғы жағындағы авторға немесе тақырыпқа нұсқап, тінтуірдің сол жақ батырмасын басқан кезде терезеде ашылып көрінеді.

Мәтіннен ізделген сөз айырықша түспен айқын көрініп тұрады және оның лингвистикалық белгіленімі меңзерді сөзге апарып сол жақ батырманы басқанда шыққан терезеде көрінеді.

Әрбір лингвистикалық белгіленім кезінде тілдік категориялар, тұлғалардың атаулары қысқартылып берілді. Қысқартулар шартты. Оның толық ашылып берілуі «Қысқарған сөздер мен шартты белгілер»* парақшасында берілген.

«Қателер жайлы хабарлау» тетігі арқылы қолданушы корпуста кезіккен қателердің сипатын жазып жібере алады.

Метамәтіндік (библиографиялық) белгілеу қолданушыға іздеу аймағын белгілі бір типтегі мәтіндермен шектеуге, яғни ҚТҰК ПМК-мен жұмыс істеуге мүмкіндік береді. Метабелгіленім 12-20 параметр бойынша белгіленген.