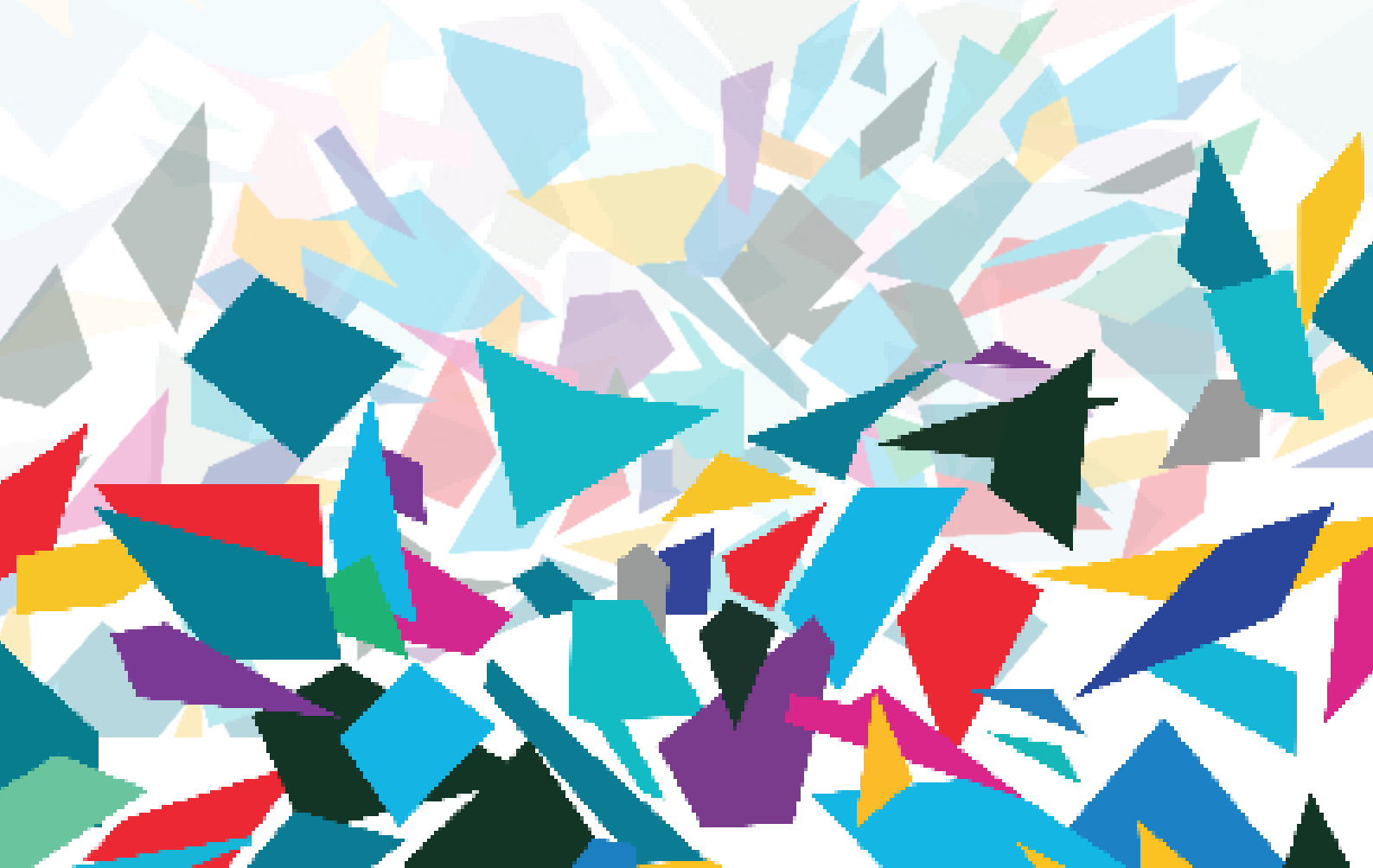
24-дәріс

ҚАЗАҚСТАННЫҢ АШЫҚ УНИВЕРСИТЕТІ

ТІЛ БІЛІМІНЕ КІРІСПЕ

Тілдің компьютерде өңделуі. 2-бөлім

6-тарау. Тілді меңгеру



Бұл дәріс компьютердің тілге қатысы жайлы. Компьютерлердің тілге қатысты кез келген салада тиімділігін айтып отыру артық болар еді. Оның біраз саласын өзіміздің компьютерлік лингвистика тармақтарын, табиғи тілдердің түрлі компьютерлік бағдарламалармен қарым-қатынасын зерттеуде қамтып өттік. Бүгін лингвист мамандар грамматикасының табиғи тілге сәйкестігін тестілеуден бастап, тілдік мәселелерді шешу және сот лингвистикасы деңгейіне дейінгі компьютерлердің қолданыс аймақтарын талдаймыз.

Лингвистердің адам тіліне арнап құрастырған грамматикасы компьютерлер қолданатын грамматикамен бірдей болмауы мүмкін; көбінесе лингвистикалық сипаттау модельдері бойынша мүлдем ұқсас болмауы да мүмкін. Адамдар мен компьютерлер әртүрлі болғандықтан, түрлі ұқсас нәтижелерге әртүрлі жолмен жетеді. Қарапайым сөзбен айтқанда, кез келген ушу құрылғысы кез келген құстың дәлме-дәл қайталамасы бола алмайды. Сол сияқты компьютер грамматикасы адам тілінің грамматикасына барлық жағынан ұқсай бермейді. Қалай дегенмен де, компьютерді адам тілі грамматикасына негізделген ережелерге сәйкес модельдеуге болады. Ең дұрыс грамматика үлгісі – адам тілінің барлық сөйлемдерін құрастыра алатын грамматика. Сөйлем құра алмау – грамматикалық тұрғыда қате сөздер тізбегі сияқты көрінуі мүмкін. Алайда табиғи тілде адам баласы грамматика бойынша қате сөз тіркестерін айта алады – үзік сөйлемдер, асығыста айтылған сөздер, спикердің сөйлеу әрекетіне қатысты түрлі ағаттықты байқасақ, қате деп баға берер едік. Оған себеп, грамматика деген – шексіз, оның лингвистикалық ережелерге сәйкестігін тексеру үшін өте көлемді жұмыс атқарылуы тиіс. Осы тұрғыда компьютерлердің маңызы зор.

Лингвистикалық грамматиканың компьютерлі моделі 1960 жылдардан, Лос Анджелестегі Калифорния университетінің синтаксис мамандары құрған ағылшын тілінің генеративті грамматикасын тестілеуге арналған бағдарламаның құрылуынан бастау алады. Сол бағдарламалар тілдегі жаңа теорияларды зерттеп, тестілеуде әлі де қолданылуда. Компьютерлік лингвистика мамандары сөйлем құра алатын және адам баласы оларды өңдеуде қазіргі лингвистикалық теорияларды қолдануға итермелейтін компьютерлік бағдарламаларды құрастыруда. Компьютерлік модельдер тілдік өнім шығаруда және егер ол кем дегенде адам тіліне жартылай болса да ұқсас сипатқа ие болғанда, оны өңдеп-түсінуде бағдарламаға енгізілген грамматиканың қолданылуы мүмкін деп есептеледі. Алайда мұндай грамматиканың расында да адамның тіл өнімдерінің өңдеу модельдеріне сай келетіні екіталай.

Джонатан Свифт адам тіліне қатысты компьютер арқылы орындалатын жиілікті өте дұрыс болжап айтқан болатын. Ол – статистикалық анализ. Ол бойынша, әріптер мен дыбыстардың, морфемалардың, сөздердің, сөздік категориялардың, тіркес түрлерінің және тағы басқаларының салыстырмалы жиілігін кез келген корпусқа жылдам әрі тиянақты түрде енгізуге болады.

Жазбаша американдық ағылшын тілінің бір миллион сөзінің жиілік анализі оның ең жиі кездесетін the, of, and, to, a, in, that, is, was, және he сөздері секілді он миллион сөзін анықтай алады. Осы сияқты «маңызды емес» сөздер тілдік корпустың 25%-ына тең, ал the жеке өзі ең жоғарғы көрсеткішке ие яғни, 7%. Ал ауызша американдық ағылшын анализі басқа нәтиже көрсетеді. Мұнда жиі қолданылатын корпустың «жеңімпаздарының» 30%-ы I, and, the, to, that, you, it, of, a және know сөздері болды. То көмекші сөзінен басқа ағылшын тілінің барлық көмекші сөздері ауызшаға қарағанда жазбаша тілде жиі кездесті және жекіру, табу сөздер саны жазба тілге қарағанда ауызша тілде көп болды.

Жиілік анализі қолданыста бар мәтіндер негізінде талдау арқылы авторлардың стилін анықтай алды. Мысалы, Інжілдің түрлі кітаптарын талдай отырып, қандай автордың қай бөлікті жазғанын анықтау мүмкін болды. Федералистік Қағаздар еңбегін зерттегенде, авторы анықталмаған парақтар Александр Хамильтонға емес Джеймс Медисонға тиесілі екені анықталды. Бұл нәтижеге қол жеткізу үшін екі жазушының белгілі жұмыстары авторы белгісіз парақтармен статистикалық тұрғыдан да сараланды.

Сәйкестік мәтін ішіндегі әрбір сөзді және оның мәнмәтінін анықтау арқылы жиілік талдауын одан әрі дамытады. Алдыңғы параграф пен кейінгі параграфтың сәйкестігін көрсетуде сәйкестік ұғымы words сөзі бес рет кездесетінін айтып қана қоймай, оның параграфтың қай жолында екенін де көрсете алар еді. Егер кімде-кім «window» сөзінің мәнмәтінін көрсетуде екі жағынан да үш сөзден таңдайтын болса, words сөзіне сәйкестілік келесі кестедегідей болады:

of one million	words	of written American
most frequently occurring	words	the, of, and,
These “little”	words	accounted for about
percent of the	words	in the corpus,
profane and taboo	words	.

Табиғатына орай сәйкестіктің қолданысы шектеулі болуы мүмкін. Сәйкестікті жақсартуға сөз тіркестері арқылы қолжеткізе аламыз. Сөз тіркесі деп екі немесе одан да көп сөздің тілдік корпуста қысқа қашықтықта қатарласып келуін айтамыз. Оның мақсаты – мәтіндегі бір сөздің кездесуі келесі сөздің кездесуіне келіп соқтыратынына дәлел табу. Сенімді нәтижелерге жету мақсатында мұндай талдауда статистика болуы қажет және зерттеу үлгісі көлемді болуы тиіс. Жоғарыда айтылған words сөзінің сәйкестік талдауында берілген мәліметтер жеткіліксіз. Егер толық бір кітаптың сәйкестік талдауын жасар болсақ, words пен written, words пен taboo және words пен of тіркестері words пен million тіркесіне қарағанда жиі кездесетінін көрсететін белгілерді анықтар едік.

Дыбыс сәйкестігі адам баласы анықтауы мүмкін емес поэзиядағы дыбыстық белгілерді анықтай алады. Компьютерлердің қолданысы әдебиет мамандарына ассонанс, аллитерация, метр және ритм сияқты поэтикалық және просодиялық ерекшеліктерді зерттеуге мүмкіндік береді. Қазірде компьютерлер қағаз бен қаламның ұқыпты жұмысын талап ететін, жалықтыруға ұшыратын көптеген механикалық жұмысты бір мезетте орындай алады.

Сэмюэл Джонсонның алғашқы сөздіктерінен бастап қазіргі жетілдірілген Oxford English Dictionary (OED) стандартты сөздіктер – компьютерлік лингвистика мамандары үшін қолайсыз. Себебі компьютерлік лингвистика мамандары үшін компьютерлік түсіну, табиғи тіл генерациясы, машиналық аударма және басқа да операцияларды жүзеге асыру мақсатында жеке сөздер мен морфемалар туралы ақпараттың бай болғаны керек. Олай болса, компьютерлік лексикография саласы стандартты сөздіктермен қатар компьютерлік лингвистикаға арнайы жасалған электронды сөздіктерді құрастырумен де айналысады.

Компьютерлік лингвистика мамандары қажет ететін ақпараттар тізімі мыналар:

- Фонемалық транскрипция;
- Фонетикалық варианттар (диалектік, әлеуметтік);
- Силлабификация;
- Синтаксистік категориялар яғни, абстарктылы, жанды, жансыз және басқа семантикалық ұғымдар;
- нөмір ұғымы, мысалы, people көпше түрде, person жеке түрде;
- Гендер ұғымы, мысалға ship аналық тек;
- c-таңдау ұғымы (murder сөзі тура толықтауышты талап етеді);
- s-таңдау (murder сөзі субъектінің де, объектінің де адам болуын талап етеді);
- стилистикалық деңгей, (мысалы, ain’t бейформальді сөз, rad слэнг, және т.б.);
- синонимдер, антонимдер, мүмкін болар омофондар, т.б.

Wordnet – семантикалық қатынасқа негізделетін, компьютер мамандарының қажетін өтей алатын он мыңдаған ықтимал жолы бар әрекетті қамтамасыз ете алатын онлайн сөздік. Осыған ұқсас басқа да еуропалық тілдердің EuroWordNet сияқты, Балқан тілдерінің BalkaNet сияқты және көптеген басқа да тілдік қауымдастықтың жобалары жасалу үстінде. Бір кездері ағылшын тілінің бір миллион жазбаша сөздік қоры «есте қаларлық» жаңалық болған еді.

Қазіргі таңда ағылшын тілінің 361,000 сөзінің орнына 361 миллиард сөзі қорға еніп отыр. Бұл қорға енген сөздер ағылшын тілінде жарық көрген бар кітаптың 4%-ын құрайтын 5 миллион кітаптан алынған. Осы сөздік қорды саралайтын арнайы зерттеуге «культуромика» деген атау беріліп, бірқатар қоғамдық ғылымда лексикография және грамматика салаларындағы өзгерістерге айрықша мән беріліп отыр. Белгілі бір өлшем қалыптастырып қарасақ, 1500 жылдан бері миллиард сөзде бір рет болса да кездесіп отырған әріптер ағылшын сөзінде болған. Олардың «жарты миллионы» – белгісіз сөздер яғни, slenthem «музыкалық аспаптың бір түрі» сияқты сөздер ағылшын лексикасында бар деп есептеледі. Шындығында, ағылшын сөздік қорының 52%-ын құрайтын ағылшын кітаптарында қолданылған сөздердің көбі стандарт сөздіктерде тіркелмеген «мағынасы жоқ» сөздерден тұрады.

Культуромикалық төңкеріс бұрыс етістіктердің өзгеріске ұшырағаны жөнінде тарихи нәтижелер көрсетті. Өткен шақ формасындағы бұрыс етістіктер «-ed» жалғауы бар дұрыс етістіктермен өкшелей іргелес қолданылып келе жатыр. Культуромикалық сараптама көрсеткендей, 1800 бен 2000 жылдар арасындағы уақытта өткен шақ формасындағы алты бұрыс етістік burn, chide, smell, spell, spill, thrive сөздері burnt, chid, smelt, spelt, spilt, throve сияқты тұрақты формаға ие болып, өткен шақ формасындағы light/lighted, wake/waked дұрыс етістіктері light/lit, wake/woke бұрыс етістіктеріне айналды. Етістіктің тұрақтылығы дұрыс етістік формасының қолданылу аясына байланысты анықталады, мәселен, chided етістігінің қолданылуы 1800 жылы 10% болса, 2000 жылы 90% болған. Тіпті өзгеріске ұшырау жиілігі компьютерлік сөздік қорда тіркелген, chide – өте жылдам өзгеріске ұшыраған етістік, spill тұрақты жылдамдықта орналасқан, бірақ, thrive бейімделуі мен басталуына қарай өзгеріске ұшыраған. Қазіргі таңда бұрыс етістік формасына қарай жылдам жылжып бара жатқан етістік sneak, ағылшын тілінде сөйлейтін халықтың 1%-ы sneaked формасының орнына snuck формасын қолдануға бейім.

Лингвистердің тарихи жазбаларына деген қызығушылық пен сөздік қорға сай t-жалғауымен аяқталатын мысалы burnt, smelt, spelt, spilt сияқты бұрыс етістіктерді зерттеуі, Америка Құрама Штаттарынан бастау алады. T-жалғауымен аяқталатын бұрыс етістіктер әлі де британдық ағылшын тілінде қолданыста бар, дегенмен жиі емес. Жыл сайын саны жағынан Кембридж көлеміне сай келерлік халық burnt формасының орнына burned формасын қолданады. Осы жазба сияқты біз лингвистикалық мүмкіндіктері мол сөздік қорды зерттеу арқылы сырттай ғана қарастырып отырмыз. Бүгінгі таңда барлық кітаптың 12%-ы компьютерге енгізілген, бұл процесс жалғасын табуда. Бұдан басқа өзге қағаз басылымдар сөздік қорға әлі енгізіле қойған жоқ. Корпустық сараптаманы зерттеуге байланысты әлемде тамаша жаңалықтар ашылуда.

Әлеуметтік ғылымдарда твиттерология awesome және fantastic сияқты жағымды эмоционалды сөздердің немесе panic және fear сияқты жағымсыз сөздердің қолданылу жиілігін тіркеу арқылы эмоционалды сөздерді қолдана отырып, ойын жеткізу сараптамаларында қолданылады. Ол еліміздегі болып жатқан күнделікті өмір салтына бастайтын жол.

Орталық барлау басқармасы сияқты ұйымдар Осама бен Ладеннің өлімі сияқты ірі оқиғалардан кейінгі аудан/қала/ел ахуалын бағалау мақсатында күніне 5 миллионнан астам Твиттер жазбасын сараптайды және үндеу жазбаларын, мемлекеттік мекемелердің билікке қарсылық көрсету немесе зорлық-зомбылық көрсеткен іс-шараларға қарсы тұруда өзінің ойын, саяси білімін ашуға көмек беру үшін айтқан мағынасыз слэнгтер мен жаргон сөздерді де жинайды.

Лингвистикалық мақсатта диалект сөздер әр ауданда және демографиялық топтарда қалай ерекшеленетінін түсіну үшін Карнеги – Меллон университетінің лингвистері твиттерологияның көмегімен зерттеу жүргізіп жатыр. Бұл зерттеулер бай сөздік қорға байланысты Солтүстік Калифорнияда «It's hella cold out there» сөйлеміндегі бір нәрсені атап көрсету үшін қолданылатын форма ретінде hella диалект сөзі жиі қолданылатынын байқау қиын екенін анықтап отыр.

Түрлі фонетикалық дыбысталулар әртүрлі нақты формаларды анықтай алады. Мысалы, Нью-Йорк тұрғыны something сөзінің орнына suttin сөзін қолданады, ал калифорниялықтар cool сөзінің орнына koo немесе coo сөзін жазады. Эмограммалар да бір аудан мен екінші аудан аралығында, бір әлеуметтік топ пен екінші әлеуметтік топ арасында түрліше қолданылады. Көп адам интернет желісін ақпарат алу үшін қолданады. Әдетте адамдар бір немесе бірнеше кілтсөзді тереді, компьютер осы сөздерге байланысты бар ақпарат көздерін немесе арнайы сайттарды тауып береді. Бұл – ақпарат алудың бір тәсілі. Кілтсөзді енгізе салып қажетті ақпаратты алу үшін интернет сайттарды табу өте жеңіл болып көрінуі мүмкін, бірақ, көбінесе заманауи лингвистикалық сараптама қолданылады. Кілтсөздердің немесе тіреқсөздің қолданылу жиілігіне, морфологиялық формаларына, мағыналас сөздеріне, семантикалық жақтан сәйкес ұғымдарға байланысты интернет сайттар жауап береді. Мысалы, bird тіреқсөзіне байланысты bird, birds, to bird, bird feeders, water birds, avian, sparrow, feathers, flight, migration, basketball great Larry сөздеріне қатысты түрлі ақпарат алуға болады.

Гугл сияқты компаниялар ақпарат алуды мультимиллиондық табыс көзіне айналдырып отыр, бұл культуромиканы белсенді ететін үлкен сөздік қор жасай отырып, ақпарат алудың басқа да салаларына жол ашады. Жалпы алғанда, ақпарат алу дегеніміз – орасан зор дерек көзіне мәлімет енгізу мен оны көрсету үшін компьютерлерді қолдану деген сөз. Ақпарат алу

жүйесіне ену компьютер лингвистикалық тұрғыдан сараптап беретін сөздерден, сөйлемдерден немесе сұрақтардан тұрады, содан кейін олардың нәтижелерін өзекті ақпарат алу үшін мәлімет көздерін сүзгіден өткізу үшін қолданады. Қазіргі таңда кешенді ақпарат алу жүйелері заманауи лингвистикалық және статистикалық сараптамаларды қолдана отырып, пайдалы құрылғыларды, сөздік қормен байланысты немесе тағы басқа компьютерлік мұрағаттарды анықтайды. Мәлімет табу, білім алу, мәлімет аналитикасы сияқты термин сөздер қазіргі кезде жоғары дамыған ақпарат алу жүйелерінде қолданылады.

Егер Bird сияқты тіреқсөзбен сайттан тыңғылықты іздейтін болса, күнделікті өмірден алған ақпаратқа қарағанда он есе көп ақпарат алуға болады. Осы жазбаны жарыққа шығарған күн туралы іздеу 637 миллион рет жүзеге асқан, ол үш жыл бұрын 200 миллион, жеті жыл бұрын 122 миллион болған еді. Мәліметтердің көбі қайталанады және кейбір ақпараттар басқа ақпараттарды басып озады. Жинақтау бағдарламалары арқылы компьютерлер артық ақпаратты қысқарта алады және негізгі ақпаратты анықтайды. Дүниежүзіне танымал көшбасшылар, ұйым басшылары, тіпті университет профессорлары көлемді мәтіндік материалдарды, мәселен, баяндамалар, газеттер, ғылыми мақалаларды компьютерлендіруге ниетті. XXI ғасырдың екінші онжылдығындағы жаңалық – көптеген материалды компьютерге негіздеп оқуға лайықты формаға келтіру жинақтау процесінің арқасында жүзеге асып отыр. Әдеттегі әрекет жоспары ақпарат алуға, мысалы құстар туралы жүздеген мақаланы табуға мүмкіндік береді. Әрбір мақала 5000 сөзден тұруы мүмкін. Мақаланы нақты көлемге яғни, 1/10 немесе 1/100-ге қысқарта алатын жинақтау бағдарламалары қолданылуда. Адамзат баласы соңғы нұсқаны оқиды. 500000 сөз 5000 немесе 10000-ға дейін сөзден тұратын, он немесе жиырма минутта оқи салатын ең маңызды ақпаратты ғана қамтып қысқартады.

Сөздің дұрыс жазылуын және болашақта көрініс тауып қалатын сөздің дыбысталуын тексеретін құрылғылар – ақылға қонымсыз, сөздіктен сөздерді үздіксіз қарай берудің орнына you're сөзінің орнына your сөзі, bare сөзінің орнына bear қолданылуы тиіс деп «ақыл қосатын» өзгеше компьютерлік лингвистиканың қосымшасы. Адамдар әрқашан сөздің жазылуын тексеретін ақпарат алу жүйелеріне келіп тірелетін, іздеуге кедергі келтірмеу үшін сөздердің қате жазылуының алдын алу үшін тіреқсөздерді тексеретін құрылғыларды жасай алады. Көптеген электронды жүйе сөздің жазылуын тексереді, дегенмен бұл сипат стандартқа сай емес мәтін құрғанда қажет болмауы мүмкін. Дегенмен төмендегі өлең жолдары көрсеткендей, сөздің жазылуын тексеретін құрылғылар дұрыс/тыңғылықты түзетудің орнын баса алмайды:

I have a spelling checker,
It came with my PC.
It plane lee marks four my revue
Miss steaks aye can knot sea.
A checker is a bless sing,
It freeze yew lodes of thyme.
It helps me right awl stiles to reed,
And aides me when aye rime.
To rite with care is quite a feet
Of which won should bee proud,
And wee mussed dew the best wee can,
Sew flaws are knot aloud.

Тілдер арасында аударма жасау қажеттігі қазіргі жаһандық қоғамдағыдай ешқашан өзекті болған емес, үлкен көлем мен мәселенің күрделілігі аударма барысында компьютер көмегін қажет етеді. Табиғи тілдер процесінде алғаш рет компьютерлерді қолдану 1940 жылдары автоматты машиналық аударманы дамытуға қадам жасаудан басталды. Ол кезде аударма таңбаларды түсіндіріп беруге ғана қажет болатын. Автоматты аударманың мақсаты – түпнұсқа тілдегі ауызша мағынаны немесе жазбаша ойды енгізу және аударылып отырған тілде грамматикалық жақтан дұрыс соған ұқсас ойды қабылдау. Машиналық аударманы алғаш енгізгенде оны компьютердің жадына түпнұсқа тілдің және аударылып отырған тілдің сәйкес морфемалары мен сөздерін енгізу арқылы жүзеге асатын процесс деп сенген. Аударма бағдарламасы берілген сөйлемдердің морфемаларын аударылып отырған тілмен

сәйкестендіруге әрекет жасады. Өкінішке қарай, машиналық аудармамен жасалған ертерек-тегі тәжірибелер көп сәтсіздікке ұшырады.

Аударма – сөзді сөзбен алмастыра салу емес, ол күрделі процесс. Аударылып отырған тілде сөздің баламасы болмауы мүмкін. Ағылшын тіліндегі the red house испан тіліндегі la casa roja сияқты сөздердің орын тәртібі әртүрлі болуы жиі кездеседі. Сондай-ақ идиомаларды, метафоралар мен слэнгтерді, тағы соған ұқсас сөздерді аударуда қиындықтар туындайды. Машиналық аудармаға қарағанда, аудармашылар бұл қиындықтардан оңай жол таба алады, өйткені, олар екі тілдің грамматикасын меңгерген және тақырыппен, бүкіл әлемге белгілі сөздің жақын мағыналарымен жақсы таныс. Тіпті олардың аудармалары сәтті шықпай жатса да, ағылшын тілінде сөйлемейтін елдерде саяхатшыларға төмендегідей «көмекші құралдар» ретінде басылым белгілері шығарылады:

Лифт ертең дайын болады. Жасалғанға дейін сіздерден кешірім сұраймыз (Бухарест қонақүйіндегі жазу).

Діни қызметкерлер діни сенімге қарамастан, барлық ауруларға көмек қолын ұсынады (Швейцариялық діни ауруxана).

Су тек менеджердің рұқсатымен беріледі (Германия қонақүйі).

Жатын бөлмеде көңіл көтеру үшін басқа жынысты қонақ қабылдауға болмайды, бұл мақсатқа қонақүйдің кіреберісі пайдаланылғаны дұрыс (Цюрихтегі қонақүй).

Мұндай «аудармалар» дұрыс сөз таңдай білуде қиындық туғызады, бірақ, машиналық аудармада дұрыс сөзді таңдай білу – жалғыз түйткіл емес.

Синтаксистік мәселе де қиындықтар туғызады. Ағылшын тіліндегі ілік септігінің жалғауларына байланысты синтаксистік қиындықтар да бар, мәселен, that man's son's dog's food dish немесе the guy that my roommate is dating's cousin тіркестері. Осындай сөйлемдерді мағынасын еш өзгертпей, мұндай құрылымдары жоқ басқа тілдерге аудару сөйлем құрылымын толығымен өзгертуді қажет етеді.

Африкаанс тілінен бастап, еврей тіліне дейін ондаған тілде жеке сөзді интернет желісі арқылы аударуға болады, бірақ, адамдар ұзақ сөйлемдерді, мәселен газет мақалаларын дұрыс аударуды қажет етеді. Кәсіби аудармашылар өз аудармаларын қолдау үшін көп уақыт пен күш жұмсап, компьютерлерді тіпті жиі қолданады, бірақ, олар грамматикалық және мағыналық жақтарын ескеріп, тура аударма беруге тырысады. Біз жазбаша мәтіндер аудармасын үстірт қана қарастырып отырмыз. Бір тілден екінші тілге ауызша аудару қалай болмақ? Бір жағынан ауызша сөйлеуді тану үшін «сөйлеуді мәтінге айналдырамыз». Екінші жағынан, «сөйлеу үшін мәтін түзу» қажет болады. Жалпы, машиналық аударма осы дәрісте талқыланған компьютерлік лингвистиканың барлық салаларын қысқаша баяндап береді. Біріккен Ұлттар Ұйымы сияқты халықаралық ұйымдардың отырыстарында ауызша сөйлеуді кәсіби аудармашылар деңгейінде машиналық аудару әлі де жолға қойылған жоқ.

Сот лингвистикасы – құқық және сот салаларында қолданылатын тілге қатысты лин-гвистиканың бір саласы. Ол авторлық құқық, құқықтық тілді интерпретациялау, тілдік құқық және оны сот залында қолдану, арыздарды талдау, мысалы, өз-өзіне қол жұмсау жазбаларын, тауар белгісін бұзу, спикерді анықтау, мәтіндік аутентификация, мысалы, плагиат мәселесі, ерін арқылы оқудың заңдылығы және т.б. сияқты мәселелерді қарастырады. Компьютерлік сот лингвистикасы деп – сот лингвистика мәселелері бойынша компьютерді қолдану саласын айтамыз. Бұл дәрісте оның үш қолданысын қарастырамыз: сауда белгілері, заң терминдерін талдау және сөйлеушіні анықтау.

Заң тілі мағынасының нюанстары – сот жүйелерінің бүкіл тарихында әрдайым талқыға түсіп келе жатқан мәселе. Мысалы, «Тән қадағы» ұғымының заңдық дефинициясы – Шекспирдің Венециандық көпес қойылымының негізгі сюжеті. Сондай-ақ, жақын арада болған бір жағдай виза сөзінің несие картасының сауда таңбасы ретінде емес, заңды түрде халықаралық саяхаттауға қатысты қолданылатын сөз ретінде қолданатынын анықтап берді. Мұндағы сұрақ: виза саяхаттаушыға виза беруші елге кіруге рұқсат бере ме немесе ол мүлдем басқа нәрсе ме? Компьютерлік лингвистика маманы бірнеше миллиондаған сөзден құралған Англия банкінің корпусын зерттеп, виза және визалар сөзінің беру, бас тарту, өтініш білдіру, қажет және талап ету сияқты етістіктермен байланысып келетін жетпіс төрт жағдайын анықтап, орта деңгейлі саяхатшы үшін виза дегеніміз – белгілі бір елге кіруге рұқсат түрі деп түсінетінін айқындады.

Бұл британдық соттың шешімі халықаралық құқық саласына әлі де өз әсерін беруде, алайда мәселе толықтай шешілді деп айтуға болмайды. Осы секілді талдаулар компьютер арқылы жүзеге асырылатын тәсілдерге орай туындайтын даулы мәселелерді зерттеуде тиімділігін көрсетеді. Компьютерлік сот лингвистика саласы заңды түрде дұрыс шешімдер жасау үшін дерекқорларды түрлі жолмен зерттеуді негіз етіп отыр.

Көптеген қылмысқа аноним хабарлама себеп болады. Спикерді анықтап табу деп – жоғарыда айтылған жағдайларды шешуде адамның көмегіне емес, компьютердің көмегіне сүйенуді айтамыз. Спикерді анықтап табуда екі компьютерлік құрылғы қолданылады. Бірі – сөйлеу кезінде амплитуданың өзгерісін көрсететін сөйлеудің толқындық формалары. Екіншісі – осының алдындағы дәрісімізде айтылған сөйлеу кезіндегі тоқтап қалу жиілігін көрсететін спектрограмма. Екі тәсіл де адам құлағы естімей қалуы мүмкін сәттерді көз алдымызға графикалық түрде бейнелеп береді, соның арқасында сот талқылауларында өте тиімді құрылғы бола алады.

Бірде бомба жайлы хабарлама келіп түседі. Бұл хабарламаға байланысты Солтүстік Каролинада туып-өскен афро-американдық азамат күдікті деп танылды. Қорғау жағы спикерді анықтап табумен айналысатын компьютерлік сот лингвистика маманын куәгер маман ретінде іске тартты. Ол түрлі сөйлеу сегменттерін талдай отырып, күдіктінің хабарлама қалдырушы болуы екіталай екенін анықтап қана қоймай, хабарлама қалдырушының ана тілі ағылшын тілі болмауының ықтималдығы жоғары екенін айтты.

Куәгер маманның қорытындысынан үзінді:

The word «goodbye» occurs in the bomb threat.

Caller: Inserts an epenthetic vowel so that the pronunciation is «good-abye», clearly seen in the waveform and spectrogram. No native speaker of English is likely to have this pronunciation.The caller also pronounces «bye» with a fully diphthongized /ai/ – the way foreigners are taught.

Suspect: His «goodbye» is «goohbah», withouth the /d/ and certainly without the epenthetic vowel. His «bye» is monophtongized and somewhat lengthened as in much speech of the south, black and white.

Маман үлгілерді күдіктіден алды да, мысалда көрсетілгендей, толқын формаларын белгіледі. Анық шыққан /d/ хабарланушыда 0.40 секундтан басталып, 0.80 миллисекундтағы ұяң дыбыстың жабылуымен анықталды. Амплитудасы кішкене ғана, бірақ, шуға байланысты 0-ге де тең емес. Мүмкін болар дауысты дыбыс 0.52 с. пен 0.64 с. аралығында байқалады. 0.64 с. Кезінде /b/ ұяң дыбысының bye деген жерінде жабылуы байқалады. Оң жақта ешқандай /d/ дыбысы, тіпті қосымша дауысты дыбысы да байқалмайды, «gooh» 2.55 секундта /b/ ұяң дыбысының bye деген жеріндегі жабылуымен жалғасады.

Хабарлаушының «good-a-bye» сөзінің соңындағы өзгермелі бірінші және екінші форманттар дифтонгтардың барын байқатып тұр.

Күдіктінің спектрограммасында /d/ дыбысы да, болжаулы дауысты дыбыс та мүлде байқалмайды. Жалғыз тыныштық кезеңі bye деген сөзіндегі ұяң /b/ дыбысының алдында байқалады. Қорыта келгенде, bye сөзіндегі дауысты дыбыстың барлық форманттарының біркелкілігі оның монофтонг екенін дәлелдейді, хабарланушының жағдайында ол басқаша болды. Олай болса, компьютерлік лингвистика – адам және компьютердің қарым-қатынасын тудыратын, компьютерлердің тілдік процестерді қалай орындайтыны туралы білім. Сол сияқты компьютерлер ғалымдарға әдебиет пен тілдің талдауын жасауға, тілдерді аударуға, үлкен көлемді корпустардан керек мағлұматты табуға, қылмыстық және құқықтық мәселелерде көмектеседі.