23-дәріс

ҚАЗАҚСТАННЫҢ АШЫҚ УНИВЕРСИТЕТІ

ТІЛ БІЛІМІНЕ КІРІСПЕ

Тілдің компьютерде өңделуі. 1-бөлім

6-тарау. Тілді меңгеру

Аталмыш дәріс тілдің компьютерде қалай өңделетіні жайлы болмақ.

Компьютерлермен өз тілімізде адаммен сөйлесіп отырғандай тілдесе алсақ, біраз мақсатқа жетуде ол әлдеқайда тиімдірек болар еді. 2013 жылдан бері «компьютерлер» теледидарда, кинода адамзат тілінде сөйлей алатын және оны түсінуге қабілетті құрал ретінде көрініс таба бастады. Дегенмен сөз «түсінетін» мобильді құрылғылардың соңғы үлгілерінің өзі қаншалықты жетілдірілгенімен, олардың қабілеті адаммен ұзақ, еркін әңгіме құруға жетпейді.

Компьютерлік лингвистика фонетикадан бастап прагматикаға дейін, сөйлесе алудан сөйлемді тануға дейін, ауызша яғни, таңбалы пікірлерден оның жазбаша түріне дейін тіл мен компьютер арасындағы қарым-қатынасты барлық қырынан қарастырады.

Компьютерлік фонетика және фонология сөйлеуді өңдеумен айналысады. Оның басты мақсаты – ауызша сөйлемдерді түсінікті етіп мәтінге айналдыру және мәтіннен дұрыс ауызша сөйлем құрылуын қамтамасыз ету.

Компьютерлік морфология, компьютерлік синтаксис, компьютерлік семантика және компьютерлік прагматика салалары лингвистикалық өңдеудің жоғары деңгейлерімен айналысады.

Компьютерлік фонетика мен фонологияның негізгі екі бағыты – тілді түсіну және тілді синтездеу. Тілді түсіну – дыбыстық белгінің оның құрамдас бөліктері дыбыс пен фонемаларға жіктелуін талдау, демек, сөздің фонетикалық транскрипциясын құру. Әрі қарай өңдеу процесі жаңағы транскрипцияны экран бетіне шығару үшін немесе сөз ұғатын қосымшада қолдану үшін қарапайым мәтінге айналдырады. Ескерту: Тілді түсіну немесе сөзді тану – көпшілік ойлағандай сөйлеуді ұғынумен бірдей ұғым емес. Сөз тану – күрделі болып есептелетін ұғыну процесіне қажетті бастау.

Тілді синтездеу – берілген мәтіннің дыбыстық және просодиялық ерекшеліктерін келтіре отырып, оны сөздер мен тіркестерге біріктіру арқылы электронды белгілерді құру процесі. Бұл процесс берілген сөйлемді электронды динамикке шығару үшін немесе табиғи тілде сөйлейтін қосымшаны құру үшін қажет.

Өзге тілде сөйлеп тұрған адамды тыңдағанда, сөйлеу барысы үздіксіз екенін және кідірісті тек қана тыныс алу кезінде жасайтынын, сондықтан оның сөйлемдерін дыбыстар мен сөздерге бөлу қиын екенін байқайсыз. Сөздердің барлығы бірігіп кеткендей. Алдыңғы дәрісте біз мұны сегменттеу мәселесі деп қарастырғанбыз. Тілді түсіну барысында компьютер де дәл осындай жағдайға тап болады.

Бұрынырақта сөз не сөйлем танушы құрылғылар дыбыстарды жеке-жеке «естуге» бейімделмеген еді. Компьютерлерге жалпы тұтас сөздер, тіпті сөз тіркестерінің акустикалық бөліктерін жадында сақтап тұруы үшін арнайы бағдарламалар енгізілген, содан кейін оларға тануға тиіс кез келген ілеспе сөздерде осы бөліктерді іздеуге нұсқаулықтар беріледі. Ол кезде компьютер жадында сөздер өте аз еді, белгілі бір сөздік қор ғана болатын. Сондай-ақ ол кездегі компьютерлер түпнұсқа сөзді дыбыстаған адамның сөзін жақсы тани алған. Басқа спикердің сөзін тануға келгенде компьютерлерде ондай бағдарлама енгізілмегендіктен, кедергілерге тап болатын, ал егер айтылған сөйлем немесе сөз оның сөздік қорында болмаса, ол сөздерді тани да алмайтын. Сөздер іркіліссіз бірігіп айтылған жағдайда, оларды түсіну нақтылығы төмендейтін, ал егер сөз толықтай айтылмаса, мәселен, Mississippi сөзінің орнына missipi қолданылса, онда сөз тану сәтсіздікке ұшырайтын. Коартикуляция әсері жағдайды одан бетер ушықтыратын. Компьютер жадында his сөзі [hɪz] деп оқылады деп сақталуы мүмкін, бірақ, his soap – [hɪssop] сөз тіркесінде, his [s] – қатаң дыбысы арқылы дыбысталатындықтан, компьютер оны мүлдем оқымай қалуы ықтимал еді. Оған қоса, компьютер жадына енгізілген сөздер фонетикалық дыбысталуы жағынан ұқсас емес сөздермен толықтырылған, мәселен pilot және pirate сөздерін шатастыруы мүмкін.

Қазіргі таңда көптеген интерактивті телефон жүйелерінде сөйлемді танитын құрылғы бар. Олар сізге «1-ді басыңыз немесе «иә» деп жауап беріңіз» немесе «2-ні басыңыз» не «жоқ» деп жауап беріңіз» деуі, кейде алдын ала бір немесе бірнеше сөзден тұратын жауап нұсқасын ұсынуы мүмкін. Заманауи ұялы телефондар қолданушыларына кеңсеге қоңырау шалу немесе күнтізбені көрсету сияқты толық сөз тіркестерін қолдануға мүмкіндік береді. Мұндай жүйелердің сөздік қоры өте аз, сондықтан негізгі сөздің акустикалық үлгісіне ұқсайтын дыбыстық белгіні тез іздеп табады және әдетте қате жібермейді.

Қазіргі таңда жеке компьютерге орнатуға болатын үздік сөз тану бағдарламаларының сөздік қоры әлдеқайда ауқымды. Барынша нақты болуы үшін мұндай компьютерлер белгілі бір адамның дауысына бейімделуі тиіс және дыбыстық белгіде жеке дыбыстарды тануы тиіс, ол білінер-білінбес болса да байқай алатындай болуы керек. Себебі жеке дыбыс үшін бүкіл бір сөйлемнің мағынасы елеулі өзгеріп кетуі мүмкін.

Ақпаратты компьютерге алдын ала мәлім бірнеше сөйлем мен мағынадан тұратын пернетақта орнату арқылы бейімдейді, ол қолданушыға тән әрбір дыбысты акустикалық бөлшектерге бөледі. Кейінірек компьютер осы бөлшектерді сөздерді тану кезінде қолданады. Бұл айтушыға байланысты сөйлемді тану деп аталады.

Екі сөйлем ешқашан бірдей дыбысталмайды және айтылымда, әдетте шу пайда болады, сөйлемді тануды атап көрсететін салыстыру процесі статистикалық болады. Фонетикалық деңгейде санау құрылғылары өзінің ақпаратында [l]-дың созылуын 65% сенімділікпен және [r]-дың созылуын 35% сенімділікпен жіктеп бере алады. Басқа факторлар осы шешімге көмек беру үшін қолданылады. Мәселен, егер компьютер сенімді болса, одан кейінгі дыбыс [d] болып сөзді бастап тұрады, содан кейін [r] артынша тұрады, өйткені, ағылшын тілінде ешқандай сөз [dl]-ден басталмайды. Компьютерлік жүйе фонотактикалық шектеулікті білуі нәтижесінде артықшылыққа ие болады. Екінші жағынан егер дыбыс сөздің басында тұратын болса, онда бұл [l] немесе [r] фонемасы екенін анықтау үшін ақпарат қажет болады. Егер келесі дыбыстар [жоғары] болса, онда [l] болады, өйткені, loop сөз болып есептеледі, бірақ roop сөз болып есептелмейді. Компьютер мұны шешуге қабілетті болмаса, ол late және rate сияқты бірнеше нұсқа ұсына алады және жүйені қолданып отырған адамнан оның шешуін сұрайды. Қолданыста бар заманауи сөз танушы құрылғылар қазіргі таңда тілдің орасан зор қорына ене алады және шешімге көмек беру мақсатында сөздердің бұрынғы кездесу жиілігін қолдана алады. Егер таңдау look пен rook-ке түссе, look жиі кездескен болар еді.

Заманауи дамыған сөйлемтанушы құрылғылар сөздерді ары қарай түрлі мағынада түсінуді болдырмау үшін синтаксистік ережелерді қолданады. Егер late/rate синтаксистік мәнмәтіні «it's too ____» болатын болса, таңдау late-ке түседі. Өйткені too зат есім немесе етістіктен кейін емес, сын есімнен кейін тұрады. «It's too late….» тіркесі «It's too rate ….» тіркесіне қарағанда жиі қолданылады. Статистикалық үлгі нақты мәнмәтінде late-тің rate-ке қарағанда көп таңдалуына компьютерді бағыттайтын фактілерге негіздей отырып құрылады.

Бұл заманауи жүйелер – компьютерлік таңдау күшіне ие бола тұра тұрақты емес. Олар кедергілерге ұшырағанда сөз тану қасиетін жоғалтады. Қолданушы көптеген коартикуляциямен жылдам сөйлейтін болса яғни, what are you-дың орнына watcha дейтін болса, көптеген шу пайда болса, сөзді тану дұрыстығы бәсеңдейді. Компьютерге қарағанда адамдар сөзді дұрыс ұғады. Спикер сөздерді түсініксіз айтса да, бәрібір оның не айтып отырғанын ұғасыз, өйткені, сізге мәнмәтін көмек береді. Шулы жерде, мәселен, отырыста артқы жақта шығып тұрған шуды елемей, билеп тұрған жұбыңызбен әңгіме-дүкен құра аласыз, өйткені, сіздің миыңыз маңызды емес дыбыстарды тазалап, бір спикердің дауысын тыңдап ұғынуға қабілетті. Мұндай нәтиже өте айқын, сондықтан оған арнайы атау да берілген, оны «коктейль-парти эффект» деп атайды. Компьютерлердің шулы ортада сөзді тануға адамдар сияқты қауқары жете бермейді, дегенмен осы зерттеу байланыс жасауға бағытталған вокалды ұялы құрылғылар мен Apple, Inc. Siri сияқты жеке арнайы жүйелердің қолданыла бастағанын көрсетіп отыр.

«Сөйлейтін құрылғы» жасау бағытындағы алғашқы қадамның бірі адам тіліне ұқсайтын дыбыс шығаратын машина құрастыру болатын.

1779 жылы осындай машинаны құрастырғаны үшін Христиан Готлиб Кратценштейн жүлде алды. Бұл – «дауысты дыбыстарды нақты дыбыстайтын музыкалық аспапқа ұқсас жасалған құрылғы» болатын.

Құрылғыны жасай отырып, Кратценштейн Ресейдің Санкт-Петербург Жоғарғы академиясы ұсынған сұраққа да жауап таба алды, сұрақ мынадай еді: «А, е, і, о, u дауысты дыбыстарының бір-бірінен ерекшеленуінің табиғаты мен сипаты қандай?».

Кратценштейн дауысты дыбыстар дыбысталған кезде ауыз қуысының пішініне ұқсас «акустикалық резонаторлар» жүйесін құрды, оны өкпеден дірілмен шығатын дыбыс органдары арқылы дыбысталуға ұқсас ауа екпіні көмегімен жасады.

Шамамен бір ғасырдан кейін сөз бен оның дыбысталуына таңғалған Александр Грейам Белл адам бас сүйегі пішінінен «сөйлейтін бас» құрастырды. Ол таңдайды, таңдай пердесін, тісті, ерінді, тілді, бетті және тағы басқа органдарды түрлі материалдармен жасап, оған кесілген резеңке бөлшектерін керу арқылы дыбыстау органдары бар темір кеңірдек орнатты. Осы элементтердің барлығын күрделі пернетақтаны бақылау жүйесі басқарады. Бұл шебер жасалған аспап дауысты дыбыстарды, ұяң дыбыстарды және тіпті кейбір тіркестерді дыбыстай алды.

Акустикалық теория мен электрониканың технологиялық дамуының арқасында сөздерді компьютерлердің көмегімен дыбыстау үлкен жетістікке жетіп отыр. Бізге сөздерді дыбыстау механизмдерінің физикалық үлгілерін жасау міндетті емес, қазір физикалық белгілерді электронды түрде шығара отырып, сөйлеу процессіне қол жеткізе аламыз.

Сөйлеу сигналдары өте кішкентай акустикалық бөлшектерге дейін қысқартыла/ыдыратыла алады. Синтетикалық сөйлеуді жасаудың бір тәсілі – дыбысталуы қажет сөз немесе сөйлемге байланысты осы құрамдас бөліктерді қажетті мөлшерде біріктіру. Бұл тәсіл қарапайым сорпа қайнатудың нұсқаулығына ұқсас, мысалы: «Екі литр су құйыңыз, оған бір пияз, үш сәбіз, картоп, бір шай қасық тұз, бір шым-шым бұрыш қосыңыз және барлығын араластырыңыз» дегендей.

Ал синтетикалық сөйлеуді қамтамасыз ететін бұл тәсіл төмендегі нұсқаулықтан тұруы мүмкін:

1. Діріл беретін дауыс желбезегі сияқты жиілігі бар дауыс ырғағынан бастаңыз. Әйел немесе бала дауысын салу үшін жоғары тонды таңдаңыз, ал ер адамның дауысын салсаңыз төмен тонды таңдаңыз.
2. Белгілі бір дауысты дыбысқа, мұрынжолды дауыссыз дыбысқа немесе созылыңқы дауыссыз дыбысқа қажетті форматтарға сәйкес келетін гармоникалық тербелістерге ерекше көңіл бөліңіз.
3. Ызың дыбыстар үшін біраз ызың қосыңыз.
4. Мұрынжолды дауыссыз дыбыстар үшін мұрын резонанстарын қосыңыз.
5. Кідіріс пен аффрикаттар жасау үшін уақытша дыбыстың дыбысталуын тоқтатыңыз.
6. Тағы басқа.

Осы нұсқаулықтардың барлығы өте түсінікті және табиғи тілге барынша ұқсас сөйлеуді қамтамасыз етеді. Ал (2) нұсқа – бұл процесте ең маңыздысы, сондықтан сөйлеу синтезінің бұл тәсілі формантты синтез деп аталады.

Көптеген синтездік айтылымның симуляция кезінде кішкентай кедергілер әсерінен болатын әлі де автоматты қасиеті немесе акценті бар, оның себебі екпіннің өзгеруі сияқты суперсегментті факторлар толығымен түсінікті болмайтынында. Егер олар дұрыс дыбысталмаса, мұндай факторлар дұрыс дыбысталмаған фонемаларға қарағанда түсінбеуге әкеліп соғады. Қазіргі таңда айтылым синтезінің басты зерттеу саласы сол тілде спикер қолданатын ұйқас пен уақыт ережелерін анықтау мен оларды бағдарламалау болып отыр. Қазіргі таңда айтылымды синтездеушілер әлі күнге дейін тар мағынада түрлі диалект сөздерді қолданып сөйлейтін адамдарды түсінуге қарағанда, күн райы болжамын оқып отырған шынайы емес/жасанды дауысты түсіну әлдеқайда қиынырақ және осы салада біраз мәселе туындайтынын айтады.

Формантты синтездің бір баламасы – конкатенативтік синтез. Конкатенативтік синтездің негізгі бірліктеріне басылып жазылған бірліктер: дыбыстар, дифондар, буындар, морфемалар, сөздер, сөз тіркестері мен сөйлемдер жатады. Дифон – сөздердің біркелкі буындалуына әсер ететін бір дыбыстың аяғы мен екінші дыбыстың басын біріктіретін тасымалдаушы бірлік. Мұндай дауыс ырғағын беретін дыбыстар жүздеп, мыңдап кездеседі. Мұндай дыбыстар адам баласы сөйлеген кезде таспаланады. Синтездеу әсері компьютермен дыбысталған үнді қалыптастыру үшін қажет жеке бірліктерді жинастырудан тұрады. Мұндай құбылыс қазіргі таңда жақсы дамыған компьютерлік күштің көмегінсіз мүмкін болмас еді.

Конкатенативтік синтездерде туындайтын мәселе адамзат тілінің жатықтығын қамтамасыз ете алу болып отыр. Бұл сөйлеу екпінін автоматы түрде соған икемдеуді қажет етеді. XXI ғасырдың екінші онжылдығында өте тамаша синтездеушілердің арқасында табиғи дауысқа жақын әсер алып отырмыз. Дегенмен мұндай дыбыс бірліктері әрқашанда табиғи қалыпта шыға

бермейді, шынайы дыбыстау қиынға соғады, сондықтан адамзат тілінде акцент жиі болады. Сөйлеуді синтездеу нәтижесін қамтамасыз ету үшін «сөйлеу үшін мәтін түзу» деп аталатын компьютерлік бағдарлама жазбаша мәтінді синтездеудің негізгі бірліктеріне айналдырады. Формантты синтездеуде сөйлеу үшін мәтін түзу процесі негізгі/ішкі мәтінді фонетикалық дыбыстауға ауыстырады.

Конкатенативтік синтездерде сөйлеу үшін мәтін түзу процесі негізгі мәтінді қажет болған жағдайда бірліктерді біріктіруге негізделген көрініске айналдырады. Буындауға негізделген синтездеуде сөйлеу үшін мәтін түзу бағдарламасы.

The number is 5557766 сөйлемін ішкі мәтін ретінде, [Өə] [nʌm] [bər] [ɪz] [faɪv] [faɪv] [faɪv] [sɛv] [ən] [sɛv] [ən] [sɪks] [sɪks] сыртқы мәтін ретінде қабылдайды.

«Синтезатор» компьютерлік бағдарламасы өз жадындағы түрлі буындарды қарастырып, шынайы ұйқас үшін жабдықталған электронды процесс үшін және буын шегараларын жұмсарту үшін оларды байланыстырады. Сөйлеу үшін мәтін түзуге және кімнің телефон шалып жатқанын хабарлауға әрекеттенген телефондар мен ұялы құрылғылар әдеттегі телефон шалушыларға түсініксіз болып қалады.

Сөйлеу үшін мәтін түзудің қиындығы жетіп артылады. Біз соның екеуіне тоқталамыз. Бірінші қиындық – жазылуы бірдей, бірақ, айтылуы басқа сөздер. Read «She has read the book» сөйлемінде [red] болып дыбысталады, бірақ, «She will read the book» сөйлемінде [ri:d] болып оқылады. Сөйлеу үшін мәтін түзу бағдарламасы қай сөйлемде берілген сөз қалай дыбысталатынын қалай біледі? Дұрыс нұсқаны беру үшін компьютердің дұрыс таңдау жасауға арналған адамзат баласы сияқты сөйлем туралы құрылымдық мағлұматы болуы керек. Құрылымнан тыс тар білім жеткіліксіз болады. Мәселен, сөйлеу үшін мәтін түзу жүйесін read сөзін алдыңғы сөз have формасында болғандықтан, [red] деп оқылуы тиіс деп шамалауымыз керек, бірақ, бұл тәсіл аса оңтайлы емес. Алдымен have сол жақтан да, оң жақтан да дыбысталады. Мысалы, Has the girl with the flaxen hair read the book? Oh, read a lot of books, has he! Осы аталған құрылымды бағдарлама тануы тиіс, атап айтсақ, has to read негізгі етістіктің көмекші етістігі болуы тиіс. Егер «read сөзін have төңірегінде [red] деп оқу қажет» деген стратегияны қолданатын болсақ және The teacher said to have the girl read the book by tomorrow деген сөйлемде [ri:d] болып оқылғаны дұрыс екенін ескерсек, қате жасаған болар едік. Әрине, мұнда read [ri:d] болып оқылуы тиіс екенін білесіз, өйткені, ағылшын тілін білесіз және ағылшын тілінің синтаксистік құрылымын түсінесіз. Екінші қиындық – ұзақ өлең жолдарының екі шумағында көрсетілген өзара байланыссыз жазылу мәселесі. Төменде берілген мысалға назар аударыңыз:

I take it you already know
Of tough and bough and cough and dough."

Ough-ке аяқталып тұрған әрбір сөз фонетикалық жақтан әртүрлі, бірақ gh [f] болып оқылуы тиіс, қай кезде оқылмауы тиіс немесе ou-ты қалай дыбыстау қажет деген ережені табу өте қиын. Заманауи компьютерлер тілдегі әрбір сөздің дыбысталуын, үлкен статистикалық анализ арқылы анықталатын олардың альтернативті дыбысталуы мен оған ұқсас оқылуын жазып алып сақтайтын қуатты сақтау қабілетіне ие болуы қажет. Мұндай сөздер тізіміне акронимдер, қысқартулар, шет тілінен енген сөздер, жалқы есімдер, бөлшектік сан есімдер және кейбір арнайы белгілер (#, &, *, %) жатады. Мұндай тізім – өте пайдалы, олар сөздердің дыбысталуын айтудан гөрі оларды есте сақтауда аса тиімді. Бұл – конкатенативтік синтездің сөздік деңгейдегі негізі. Дегенмен бұл тізім ешқашан шектелмейді. Жаңа сөздер, шет тілінен енген сөздер, жалқы есімдер, қысқартулар мен акронимдер ағылшын тілі сөздігіне тоқтаусыз ене береді және олардың дұрыс дыбысталуын болжай алмаймыз. Сөйлеу үшін мәтін түзу жүйесі сөздікте көрсетілмеген белгілер үшін өзгерту ережелерін қажет етеді. Бұлар сөйлеу барысында сөздің өзіне қарағанда кішкентай бірліктерге негізделген формантты немесе конкатенативтік синтездеушілер арқылы нәтижеге қол жеткізуі тиіс.

Тілдік синтездің қолданылу аясы өте маңызды. Ол көзі нашар көретін адамдар үшін «оқитын құрылғы» көмегімен экранда көрсетіліп, сосын дыбысталады. Қазіргі таңда бұл құрылғылар сатылымда бар. Құлағы нашар еститін немесе сөйлеу органдары нашар дамыған пациенттер синтезаторларды қолдана алады. Мәселен, Солтүстік Каролина мемлекеттік университетінің ғалымдары дыбыс шығара алмайтын, қол-аяғы тұтас сал болып қалған, тек қана бас изей

алатын склероз ауруының ауыр түрімен ауыратын адамдар үшін қарым-қатынас жүйесін жасады. «Иә» деп жауап қайтару үшін басын изеп, экранда көрсетілген сөздерді таңдайды. Ойын жеткізу үшін сөйлем құрай алады, сосын оны синтезатордың көмегімен дыбыстайды.

Компьютерлердің ағылшын тілінің грамматикасын қолданып сөйлегенін және түсінгенін қаласақ, оларға морфологияны үйретуіміз керек. Компьютерлік құрылғылар «The cat is sit on the mat» немесе «My five horse be in the barn» сөйлемдерін осылай қатемен айтпауы қажет. Егер компьютерлер ағылшын тілін түсінетін болса, sitting сөзінде sit+ing деген екі морфема бар екенін, ал spring сөзінде бір морфема, reinvent сөзінде rein+vent емес, re+invent екі морфемасы бар екенін түсінуі қажет.

Сөз құрылымын компьютермен өңдеу компьютерлік морфология деп аталады. Сөздерді дұрыс түсініп, оны грамматикалық тұрғыдан дұрыс қолдану үшін компьютер сөздер құрылымын ұғынуы қажет. Сөздерді өңдеу үшін компьютерлерге сөз түбірі мен жалғауларды тану бағдарламалары қойылған. Кейбір жағдайларда бұл процесс оп-оңай орындалады. Books сөзін book+s етіп, walking сөзін walk+ing, fondness сөзін fond+ness, unhappy сөзін un+happy деп бөліп жазу қажет. Мұндай тәсілдер өте қарапайым, сөздердің жазылуы да қолайлы. Дегенмен profundity=profound+ity, galactic=galaxy+ic, democracy= democrat+y деген сөздер бағанағы сөздерге қарағанда күрделірек.

Қазіргі таңда компьютерлердің миллиондаған сөз формасын тану мүмкіндігі бар болғандықтан, түрлі қиындықтар туындауы мүмкін. Тілде жаңа сөз пайда болғанда, мәселен бірнеше жыл бұрын пайда болған blog сөзінің blogs, blog's, blogging, blogged, reblog, blogable сияқты бірнеше біртүбірлі нұсқасы пайда болды. Алайда оның басқа да формалары, мәселен blogify, exblog, disblog сияқты нұсқалары болмауы мүмкін. Сондықтан компьютерге енгізген сөздерді әр уақытта жаңартып, толықтырып отыруға тиіспіз.

Тағы бір айта кетер жайт, сөздердің барлық формасын болжау мүмкін емес. Дегенмен Heaten сөзі сөздік қорға енбеген, бірақ, бұл сөзді естігенде біз оның «жылыту» мағынасын беретінін білеміз, осындай түйсік компьютерде де болуы тиіс. Сөзжасалым – жалпы құбылыс, сондықтан ағылшын тілінде болуы тиіс барлық күрделі сөздерді болжау мүмкін емес. Pod+cast құрылымынан podcast сөзі жасалғанда, ешбір компьютерлік жүйеде бұл сөз тіркелмеген еді.

Кез келген сөйлемді түсіну үшін оның синтаксистік құрылымын білу қажет. Dogs that chase cats chase birds сөйлемінің синтаксистік құрылымын білмеген жағдайда құстарды аңдыған иттер немесе мысықтар екені түсініксіз болар еді. Осыған сәйкес тілмен жұмыс істейтін машиналар да синтаксистік құрылымды анықтай алатын болуы қажет. Парсер – спикерлердің синтаксистік құрылымын қайталауға тырысатын компьютерлік бағдарлама. Адам баласы сияқты компьютердегі парсер де сөздер тізімін реттеу мақсатында грамматиканы қолданады. Алайда парсерлер бірнеше синтаксистік категорияға жататын сөздер кездескенде қате нәтиже беруі мүмкін. The little orange rabbit hopped сөйлемінде парсер orange сөзін зат есімге жатқызып, қате нәтиже көрсетеді. Кейін, қателік анықталғанда, парсер шешімді қайтадан қарастырып, orange сөзін анықтауыш етіп, сөйлемді қайтадан талдайды. Бұл тәсіл The old man the boats және The Russian women loved died секілді мағынасы түсініксіз, бірақ, грамматикалық тұрғыдан дұрыс сөйлемдер болғанда да қолданылады.

Осындай мағынасыздық жағдайларын реттеудің тағы бір амалы – компьютердің грамматика бойынша мүмкін деген барлық талдауын бір уақытта қатар жүргізуі. Мұндай жағдайда тек аяқталған талдаулар ғана дұрыс деп қабылданады. Бұл тәсілді The Russian women loved died сөйлемін талдауда қолданғанда, екі талдау қатар анықталады: Russian бір нұсқасында анықтауыш, екіншісінде зат есім болады. Анықтауыштық талдау The Russian women loved деген жерге дейін талдауын жалғастырады. Алайда died сөзіне жеткенде тоқтап қалады, себебі, ереже бойынша етістік тіркесінде бұл сөздің орны қате болып есептеледі. Бұл талдау died сөзін анықтай алмағандықтан, аяқталмаған болып есептеледі. Екінші шыққан талдауда, Russian women екі сөздің бірге келетіні анықталғанда that сөзі women сөзінің алдында келсе, бағыныңқы сөйлем құрылымын анықтауға талпыныс жасайды. Талдау that сөзін таппағандықтан, бағыныңқы сөйлем құрылымын анықтамайды. Талдау бағыныңқы сөйлем құрылымын анықтамағандықтан, The Russian women тіркесін зат есім тіркесі деп анықтайды.

Компьютерлердің мүмкіншілігі белгілі бір жағдайларда адам баласының мүмкіншілігінен асып түсуі мүмкін, себебі, олар семантикалық түрде аңғырт болып келеді. Айталық, Time flies like an arrow мысалын қарастырайық. Компьютерлік талдау бұл сөйлемнің мағынасын жылдам

анықтайды. Бірнеше сөз бірнеше синтаксистік категорияға жатуы мүмкін екенін ескеріңіз. Бұл мысалда ондай сөздердің саны бесеу.

Компьютерлердің адам тілін түсінуден бөлек, біз олардың алдын ала енгізілмеген жаңа сөйлемдер құра алу қабілетіне ие болғанын қалаймыз. Бұл процесс синтаксистік құрылымдарды білуді талап етеді. Кей жағдайларда мұндай бағдарламаны құрастыру өте жеңіл болуы мүмкін. Мысалы, компьютерлік бағдарлама Шекспир стиліндегідей жекіру сөздерін жасау үшін үш баған ұйымдастырады. Бірінші бағанда жай анықтауыш, екінші бағанда қос сөзді анықтауыш, үшіншісінде зат есімдерден құралған сөздер тізімі орналасады:

Жай анықтауыш
bawdy
churlish
goatish"

Қос сөзді анықтауыш
beetle-headed
clay-brained
fly-bitten"

Зат есімдер
baggage
bladder
codpiece"

Зат есімдер тіркесі қорын жасау үшін компьютерлік бағдарлама әр бағаннан кез келген тізбекте бір-бір сөзден таңдайды. Нәтижесінде жекіру сөздері лезде дайын болады: you goatish, pox-marked bladder, you lumpish, milk-livered hedge-pig. Бұдан да жеңіл бағдарлама алғашқыда жүктелген идеялар мен тұжырымдарға лексикалық ұғымдар берер еді. Тілдік өнім беру мақсатында олар сол тілдің синтаксисіне сай келетін тіркестер және сөйлемдермен салаластырылуы қажет. Компьютерлік талдауда парсинг бойынша негізгі екі тәсіл бар: жоғарыдан төмен қарай және төменнен жоғары қарай. Жоғарыдан төмен талдау бағдарламасында сөйлемнің талқылануы жоғарғы деңгейге жататын сөйлем категориясынан бастау алады. Төменгі деңгейлер тілдің синтаксистік ережелерін ескере отырып, зат есім және етістік тіркестерінен басталады да, төменге қарай анықтауыш, зат есім, етістік және сөйлемнің басқа да мүшелерімен толықтырылады.

Жоғарыдан төмен қарай анықтайтын тәсілде талдау діттелген мағынаны айту мақсатында қолданылатын лексикалық ұғымдардан басталады. Артынша оларды жоғарғы деңгейлі категорияларды құру мақсатында байланыстырады. Мұнда да адам баласы тілдік мәнді қалыптастыру үшін екі тәсілдің қайсысын таңдайтыны әлі де белгісіз.

Тілдік талдаудың, яки тілді құрастырудың қай кезде синтаксистік, қай кезде семантикалық болатыны әрдайым анық бола бермейді. Time flies like an arrow мысалын қарастырғанда, барлық синтаксистік талдауларды семантикалық талдауға салғаны дұрыс бола ма, әлде садақ оғының уақытты санауды білдіретін мағынасын алып тастау мақсатында екеуін де яғни, синтаксистік және семантикалық талдауларды бірдей уақытта жүргізгені дұрыс па? Әлде, кейбір лингвистер ұсынғандай, синтаксистік талдауға алып келетін семантикалық талдауды бірінші жасаған дұрыс па? Бұл сұрақтар психолингвистика және компьютерлік лингвистика салаларындағы өзекті тақырыптардың қатарында.

Мағынаны қалай түсіндіруге болады? Бұл – мыңдаған жылдар бойы талқыланып келе жатқан мәселе және бұл лингвистика, философия, психология, когнитивті ғылымдар мен компьютерлік ғылымда да түрлі зерттеулердің жасалуына әлі де жағдай туғызып отыр. Жалпы айтқанда, компьютерлік семантика деп компьютердегі сөздер мен морфемаларды немесе олардың комбинациясынан құралатын мағынаны айтамыз.

Компьютерлік семантика қарастыратын екі маңызды мәселесі бар. Біріншісі – тілдік ақпараттың семантикалық ұғымын қалыптастыру. Екіншісі – мағына беретін табиғи тілдік өнімді құрастыра алатын семантикалық ұғымды қалыптастыру. Мұны түсіну үшін ақауды жөндейтін техник мамандарға нұсқаулық ретінде жасалған «The Circuit Fix-It Shop» деп аталатын ауызша диалогтан үзінді қарастырайық:

КОМПЬЮТЕР: This is the Curcuit Fix-It Shop. How may I help you?
ТЕХНИК: I need to fix the RS111 circuit.
КОМПЬЮТЕР: I am familiar with that circuit. The LED is supposed to be displaying alternately blinking one and seven
ТЕХНИК: Where is the LED?
КОМПЬЮТЕР: In the middle right of the blue region is the LED.
ТЕХНИК: No light.
КОМПЬЮТЕР: What is the switch at when the LED is off?
ТЕХНИК: The switch is down.
КОМПЬЮТЕР: You put up the switch.

Circuit Fix-It Shop-тың операциясы барысында компьютер жүйе қолданушысының тілдік ақпаратты талдап, семантикалық ұғымын қалыптастыруы қажет, оған сәйкесінше жауап қату және оны жүйе қолданушысына қарапайым тілде жеткізе алуы керек. Сөйлем құрау мақсатында компьютер семантикалық ұғымға сәйкес келетін сөздерді табуға тырысады. The Circuit Fix-It Shop жүйесінде келесі кезекте нені айту керек екенін шешу керек: ажыратқыш, жүйе қолданушысы, жарық немесе басқасы ма? Ол өзінің қарастырып жатқан нысанының жағдайын сипаттауға сәйкес келетін сөздерді, жүйе пайдаланушының көзімен нысанның жағдайын сұрау, жүйе пайдаланушының сұранысын орындау, жүйе пайдаланушысына әрі қарай қандай әрекеттер жасау екенін айту қажеттігін осылардың ішінен таңдау керек болады. Егер сұрақ жүйе пайдаланушысына байланысты болса, сен есімдігі қолданылады; егер ажыратқыш жағдайы негізгі мәселе болса, the switch немесе a switch above the blue light сөздері қолданылады. Бұл әрекетте мағына бөлшектері ұйқасқанда, грамматикалық тұрғыдан дұрыс ұғым жасауда біз білетін синтаксистік ережелер іске қосылады. Сөйлеуді ұғыну деңгейіне қол жеткізу үшін компьютер өзінің семантикалық ұғымдарында берілген ақпараттағы сөздер мен олардың құрылымына сәйкес тұжырымдарды табуға тырысады. Техник маман I need to fix the RS111 circuit деп айтқанда, жүйе I сөзін қолданушы деп анықтайды, need қолданушыда жоқ нәрсені меңзейтінін және оны компьютер ұсына алатын ұғым болу қажет деп біледі. Кейін жөндеу керек екенін білгенде, белгілі бір нәрсенің қалай жұмыс істейтінін компьютер түсіндіру керек екенін білетін болады. Компьютер the RS111 circuit ұғымын өзінің семантикалық ұғымында қамтылған, белгілі бір қасиеттерге ие жүйе деп анықтайды.

Компьютер тұжырымдарды бірнеше жолмен қамтамасыз ете алады, алайда оның ешқайсысы мінсіз немесе бірі екіншісінен үстем болмайды. Барлық тәсілге ортақ бір нәрсе бар: айтуға немесе түсінуге арналған сөздер мен морфемалар лексикасы. Нақты сол ақпараттың қалай құралғаны оған сәйкес келетін қолданымдарға тәуелді болады.

Семантикалық ұғымдар сызбалары компьютерлік лингвистика негізінде жасалған. Оның бірі – сөздерді олардың тақырыптық рөлдерімен байланыстыратын желілерді қолдану.

Алдында айтып өткеніміздей, прагматика – «шынайы өмір» мен тіл жүйесінің қарым-қатынасы. Circuit Fix-It Shop жағдайында компьютер ажыратқыштың біреу екенін және бұл әлемде одан басқа ажыратқыштың жоқ екенін біледі. Сол себептен анықтауыш the мұнда өте орынды орналасқан. Егер адам a wire дегенді ескертсе, компьютер жадындағы сымдардың бірнеше екенін білгендіктен, which wire деп сұрақ қоя алады. Бұл – жеңіл компьютерлік прагматика мысалы. Егер сөйлем He saw the boy with a bicycle секілді құрылымы бойынша екі мәнді болса, компьютер әр құрылымды жеке талдауға көшеді. Семантикалық талдау кейбір ауытқу құрылымдарын ескермеуі мүмкін. He saw the boy with a telescope сөйлемі екі талдауда да мағынаға ие болады және оның негізгі мағынасын білу үшін жағдаяттық прагматикалық білім қажет болады.

Лингвист зерттеушілер адамның сөйлеуін анықтай алатындай, ым тілінде түсінетін компьютерлік алгоритм ойлап табуда. Ыммен спикерлер камера алдына тұрып микрофонға сөйлегендей ымдап сөйлейді және компьютер көру процесі арқылы алдын ала енгізілген ымдар жиынтығынан тиісті ымды байланыстыруға мүмкіндік жасайды яғни, бұл аудио процесс арқылы алдын ала енгізілген акустикалық толқындар жиынтығынан тиісті үнді байланыстыру процесіне ұқсас болып келеді. Көру процесі бұл жағдайда қол қозғалысын және пішімін, оның траекториясын және орналасуын анықтайды. Алайда бұл құруға қиын және дұрыс нәтижеге жету ықтималдығы шектелген алгоритм болып келеді.