ИБРАЕВА Н.А.

К.Тыныстанов ат. ЫМУ

ЛИНГВИСТИКАДА МАТЕМАТИКАЛЫК СТАТИСТИКАНЫ КОЛДОНУУ

XX кылымдагы лингвистиканын өтө маанилүү методологиялык принциби болуп тилди система катары кароо саналат. Ушул методго ылайык, тилди лингвистикалык изилдөө тилдик системаны түзгөн элементтерди сыпаттоо жана алардын бири-бири менен болгон катышын сыпаттоо экендиги талашсыз.

Лингвостатистикада «статистика» түшүнүгү традициялуу түрдө сандык байкоо жүргүзүүнүн математикалык аппараты гана болбостон, кептеги (тексттеги) тилдик белгилердин тигил же бул сандык көрсөткүчүнүн ишке ашуусун белгилейт.

Лингвистикада статистиканы колдонууда өз учурунда тил таануучулардын арасында дискуссия жараткан 3 суроо келип чыккан: Эмнени саноо керек? Кантип саноо керек? Эмне үчүн саноо керек?

Ар түрдүү баарлашуу чөйрөсүндө ар түрдүү жыштыктар менен мүнөздөлгөн тилдик бирдиктерди саноо керек. Коюлган маселеге карата тилдин ар түрдүү бирдиктери болушу мүмкүн – лексикалык, морфологиялык, синтаксистик жана тексттик, баарынан аз өлчөмдө фонетикалык. Бирок буларды кандай саноо керек: кээ бирин ганабы же бардыгын теңби?

Мисалы, функционалдык стилдеги тексттердин бардык массивинен, алардын ички дифференциациясынан эки жол менен кетүүгө болот. Же удаалаш тилдик бирдиктерди статистикалык жол менен изилдөөгө алып, эмпирикалык жол менен кетүүгө, же маани берүүчү сөздөрдүн сөз түркүмдөрү жана морфологиялык көрсөткүчтөрү боюнча кетүү керек. Бул жол акыры келип стилдердин лингвостатикалык түзүлүшүн толук бойдон сүрөттөй алат. Бирок функционалдык-стилистикалык жол үнөмдүү болушу мүмкүн. Ишмердүүлүктүн тигил же бул чөйрөсүн эске алуу менен, ой жүгүртүүнүн типтеринин өзгөчөлүктөрүнө жараша ж.б. базалык экстралингвистикалык факторлордун негизинде алгач интуициялык деңгээлде кыйла мүнөздүү болгон кандайдыр бир тилдик формадагы изилденүүчү тексттин жетиштүү жыштыгы тууралуу гипотезаны сунуш кылууга болот (семантикалык жагын эске алуу менен), лексикалык-грамматикалык маанидеги жана аларды алгач статистикалык анализге алуу менен, б.а., бардык катары менен эмес жана бардык тилдик бирдиктер эмес, тандалмалуу гана. Алардын ичинен стилдик спецификага, подстилдин өзгөчөлүктөрүнө ж.б. болжол менен туура келгендерин гана

алуу зарыл. Ошону менен бирге эле тилдик бирдиктердин семантикалык жагына алардын туура келүүсүнө (же туура келбөөсүнө) ой жүгүртүү мүнөзүнө жана аң-сезимдин формасына жараша тигил же же бул функционалдык стилде чагылдырылышына өзгөчө көңүл буру керек. Стилдин спецификасы тилдик системанын атайы шөкөттөлгөн каражаттарын колдонуу гана пайда болбостон, функционалдык боёкчодогу жыштыктар басымдуулук кылган стиль менен түзүлөрү байкалган.

Ал эми кантип саноо керек деген суроого математикалык статистиканын эрежелерине жараша деп жооп берүүгө болот. Бул жерден илимдин ар түрдүү тармагындагы өзгөчөлүгүн эске алуу менен, изилденип жаткан объектинин спецификасына, сөзсүз түрдө коюлган маселеге жараша колдонуларын эске алуу зарыл. Эгерде фармакологияда же медицинада салыштырма катанын чоңдугу жокко эсе боло турган болсо, лингвостатистикалык изилдөөлөрдө ката канааттандырарлык деңгээлде – 5-10% барабар болсо, кээ бир учурларда чоң болушу да мүмкүн (Головин, 1971, 56). Бирок алынган жыйынтык канчалык так болсо, ошончолук жакшы. Ушуну менен байланыштуу, өзгөчө, ишти уюштуруунун башталышында тандап-иргеп алуу тууралуу суроолор келип чыгат. Бир жагынан, текстти өтө «жыш» изилдөө талап кылынса, башка жагынан, тексттеги изилденип жаткан бирдиктин кезигүүсүнүн көптүгүн эске алуу керек (өзгөчө, этиштин колдонулушу боюнча) (Кожина, 1972; гл.II).

Эмне үчүн саноо керек деген негизги суроого тилдик каражаттардын кептин ар түрдүүлүгүндө кызмат аткарышынын мыйзам ченемдүүлүктөрүн табуу үчүн жана кептин стилинин тигил же бул объективдүү экстралингвисттик факторлордон көз карандылыгын аныктоо үчүн деп жооп берүүгө болот.

Лингвостатистикалык изилдөөлөрдө лингвисттер ыктымалдык-статистикалык методго көбүрөөк басым жасашат, анткени бул метод изилденип жаткан объект тууралуу кеңири маалымат алууга жардам берет (Гладкий, Мельчук, 1969).

Вяч.Вс.Иванов өзүнүн эмгектеринде тилди изилдөөдө ыктымалдык-статистикалык методду колдонуунун 3 реалдуу негизин келтирет. Биринчи негизи «…Сөздөрдүн, муундардын жана фонемалардын сандык катышы тилдердин классификациясын түзүүгө негиз түзөт, бул изилдөөнүн жардамы менен алардын тарыхын изилдөөгө жардам берет. …Экинчи негизи болуп, тилдин түзүлүшүндөгү сапаттык жана сандык катыштын негизиндеги ички көз карандылык. …Фонемалардын саны сөздөрдүн жана морфемалардын сапатына таасирин тийгизет. Морфемалардын саны анын сапатына таасир тийгизет (тыбыш гана катары эмес, структуралык-семантикалык жактан да). Морфологиялык деңгээлдеги сандык мүнөздөмө синтаксистик кубулуштарга да таасир тийгизет. Үчүнчү реалдуу негиз, кептин жүрүшүндө тилдик элементтердин жыштыгы стаститиканын тигил же бул закондоруна баш иет» (Головин, 1971).

Ар бир статистикалык изилдөө объектилердин көптүгүнүн үстүнөн байкоо жүргүзүү болуп саналат. Бул объектилер көптөгөн белгилери менен мүнөздөлөт жана бир объектиден кийинки объектиге өтүүдө өзгөрөт. Бардык белгилерин бир эле мезгилде кароо мүмүкүн эмес, ошондуктан тилчи белгилүү бир гана белгиге терең көңүл буруусу зарыл, анткени башка белгилерине карата берилген лингвистикалык көптүк бирдей укукта. Мындай жол менен берилген көптүктү бир тектүү деп эсептөөгө болот. Көрсөтүлгөн ыкма менен түзүлгөн көптүк статистикалык жыйынды, аны түзүүчү объектилер жыйындынын бирдиктери деп аталат (Крапник, 2004).

Лингвистикалык объектилер сандык да, сапаттык да касиеттерге ээ болушат. Сандык касиеттер (мисалы, сөз формасынын тамга же фонема, мүчө, морфема түрүндөгү узундугу, же сүйлөмдөгү сөздүн колдонулушунун саны ж.б.) дайыма лингвистикалык объектилер статистикалык жыйындынын бирдиги катары колдонулат.

Бирок тексттин статистикасы сандык гана эмес, сапаттык белгилерге да байланыштуу болот. Мисалы, статистикалык-морфологиялык изилдөөнүн жүрүшүндө тексттеги сөз колдонуулар анын кандайдыр бир сүйлөм мүчөсүнө гана тиешелүүлүк белгиси боюнча топтолот. Ал эми статистикалык-синтаксистик изилдөөдө сапаттык белги болуп ар бир сөз колдонулушунун сүйлөмдүн белгилүү бир мүчөсү ролунда кызмат кылуусу эсептелет (Адмони, 1970).

Ар бир лингвистикалык изилдөөнүн ийгилиги статистикалык байкоонун уюштурулушунан көз каранды, биринчиден, лингвистикалык белгини тандоо жана жыйындынын бирдигин аныктоо, экинчиден, байкоонун жолун аныктоо.

Ар бир жыйындынын бирдигин белгилөөчү сандык жана сапаттык белги лингвистикалык түшүнүккө жана тил таануу изилдөөлөрүнүн максаттарына жооп бериши керек.

Кептин коммуникативдик касиеттеринин бири анын баалуулугунда. Бул категорияны изилдөөдө 2 аспект бири-биринен айырмаланып турат. Бир жагынан, кептин мазмундуулугунун баалуулугу, б.а., маалыматтык-логикалык, семанти-калык аспектисинин анализи. Экинчи жагынан, кептин сырткы формасын, б.а., кепте колдонулган тилдик бирдиктердин мүнөздөмөсүн баалоонун аспектиси. Текстти лингвостатистикалык изилдөөдө дал ушул аспектиге көңүл бурулат.

«Баалуу кеп» деген түшүнүктүн негизинде кеп канчалык «бай» же «баалуу» болсо, анда ошол белгилер жана белгилердин чынжырлары, кептин ички уюштурулушунун ырааттуулугу ошончолук сейрек кезигет. Бул кеп канчалык «бай» болсо, өзүнүн бардык тилдик деңгээлинде ошончолук ар түрдүү экендигин көрсөтөт. Мындай шарт кептин «байлыгынын» жана «ар түрдүүлүгүнүн» теңдеш экендигин айгинелейт.

Тексттеги тигил же бул фрагменттин чегиндеги, же толук текст

боюнча кепти баалоонун методдору изилдөөчүнүн интуициясына таянат да, жыйынтыгында мындай баалоо субъективизмге алып келет.

Тилдин баалуулугун изилдөөдө, объективдик–сандык (лингвостатистикалык) изилдөөлөрдү колдонуу сейрек кезигет. Кээ бир учурда мындай изилдөөлөр ар түрдүү тилдик деңгээлдерде, комплекстик жалпылоосу жок, практикалык эмес декларативдик мүнөздө гана бааланат.

Бул жетишпестикти кептик ар түрдүүлүктүн так жана абдан түшүнүктүү статистикалык коэффициенттерин колдонуу менен толуктоого болот.

Кептин ар түрдүүлүк коэффициенттери.

Кептин ар түрдүүлүгүнүн коэффициентин эсептөөдө эң аз дегенде эки параметрин эске алуу зарыл: лексикалык ар түрдүүлүк жана синтаксистик татаалдыктын даражасы. Бул эки деңгээлде коэффициенттердин нормалдуу катышын баалоо оңой. Анткени коэффициент абсолюттук эмес, салыштырмалуу чоңдук (чоңдуктун аныкталган аралыгында) болгондуктан, салыштырылып жаткан тексттердин белгилүү чегин көңүлгө албай деле коюуга болот. Бул жерде негизги теориялык кызыкчылык тексттин ички «динамикасын» изилдөөдө тексттин ар түрдүү бөлүгүндө (участок) коэффициенттерди салыштыруу жана бардык текст үчүн коэффициенттик катышын кароо болуп саналат.

Лексикалык ар түрдүүлүк.

Кептин ар түрдүүлүгүнүн коэффициенти лексемалардын санынын тексттеги сөздүн санына болгон катышынан түзүлөт, б.а.;

$$K_{лекс} = \frac{Л}{С} \qquad (1)$$

$K_{лекс}$ - лексикалык ар түрдүүлүктүн коэффициенти;

Л – берилген тексттеги лексемалардын саны;

С – тексттеги жалпы сөздөрдүн саны.

Коэффициенттин мааниси 0 дон 1ге чейинки аралыкта жайгашат. Алынган жыйынтыктагы мааниде ондук бөлчөк канчалык чоң болсо, лескикалык ар түрдүүлүк ошончолук жогору болот.

1-мисал. Төмөндөгү тексттин лексикалык ар түрдүүлүгүн эсептейли.

Данияр дагы көпкө чейин обон салып, ырдап келди. Чалкыган август түнү аны тыңшап, обонго арбалгандай терең тынчтыкта уюп магдырады, ал түгүл аттар да, арабаларды жай терметип, жай басышты. Мына ушинтип, адамдын денесин балкыта, толукшуп ырдап келе жаткан Данияр бир убакытта үнүн бийик закымдата келип, ырын чорт токтотту да, аттарга ала-сала камчы уруп, айдап жөнөдү. Менин оюмда Жамийла дагы анын артынан түшөт экен деп, тизгиндерди жыйып, алардын соңунан калбай утурламак болдум, бирок Жамийла козголуп да койгон жок. Башын бир жагына кыйшайткан калыбында терең кыялга чөгүп, нес болгон сыяктуу ошол бойдон былк этпей отура берди. Обондун абада калган

учкундары дагы эле көкөлөп жүргөндөй. тыъшагансыйт. (Ч.Айтматов, Жамийла аны сыйкырланып Жамийла, повесть, 1958).

№	лексема	№	лексема	№	лексема	№	лексема
1	Данияр	21	ат	41	ала-сала	61	кой
2	дагы	22	араба	42	камчы	62	жок
3	көп	23	жай	43	ур	63	баш
4	чейин	24	терме	44	айда	64	жак
5	обон	25	бас	45	жөнө	65	кыйшай
6	сал	26	мына	46	мен	66	калып
7	ырда	27	ушинт	47	ой	67	терең
8	кел	28	адам	48	Жамийла	68	кыял
9	чалкы	29	дене	49	арт	69	чөк,
10	август	30	балкы	50	түш	70	нес
11	түн	31	толукшу	51	экен	71	сыяктуу
12	ал	32	жат	52	де	72	ошол
13	тыңша	33	бир	53	тизгин	73	бойдон
14	арба	34	убакыт	54	жый	74	былк
15	терең	35	үн	55	соң	75	этпе
16	тынч	36	бийик	56	кал	76	отур
17	ую	37	закымда	57	утур	77	бер
18	магдыра	38	чорт	58	бол	78	аба
19	ал	39	токто	59	бирок	79	кал
20	түгүл	40	да	60	козго	80	учкун

81	эле
82	көк
83	жүр
84	сыйкыр

Көрүнүп тургандай, Ч.Айтматовдун «Жамийла» повестинен алынган контекстте лексикалык ар түрдүүлүктүн үлүшү жогору: 7 сүйлөм 105 сөздөн туруп, 84 лексема колдонулган.

Синтаксистик ар түрдүүлүк.

Синтаксистик коэффициенттин ар түрдүүлүгү берилген тексттеги сүйлөмдүн санынын тексттеги сөздөрдүн санына болгон катышынан келип чыгат:

$$К_{лекс} = \frac{Л}{С} = \frac{84}{105} \approx 0{,}809523 \qquad (2)$$

$К_{синт}$ -татаалдык коэффициенти;

Л – сүйлөмдүн саны;

С – тексттеги сөздөрдүн саны.

Аралык маанилер (1) формуладагыдай эле 0 дөн 1 ге чейин болот жана бөлчөктүн мааниси канчалык чоң болсо, берилген тексттеги сүйлөмдөр ошончолук көп сандагы сөздөрдөн турат, демек, бир сүйлөмдүн курамында сөздөрдүн арасында синтаксистик ар түрдүүлүктүн катышы да жогору.

2-мисал. Берилген текст үчүн синтаксистик ар түрдүүлүктү эсептейли.

Кеч кирип келе жатканы менен, күндүн аптабы али кайта элек, ал аз келгенсип, батыш тараптагы казактын чөл талаасынан кимдир бирөө көрүк желпип жаткансып, керимселдин дем кыстырган оор толкундары лап-лап этип келе баштады. Азыр айланадагынын баары: алы кеткен самсаалаган бутактар, ичи түнт токой сыяктуу жылуу ным басып, чарбактарда демигип турган чатырман жүгөрүлөр, чымынга таланып арык боюнча ылаалап турган боз бээ, мына ушунун бардыгы жайкы ысыкка чагылып, азыр түндүн келишин гана тилеп, күтүп тургансыйт. Түндө жогортон муздак жел согот, таңга жуук жер үстүн шүүдүрүм камайт, ошондо таш "чыйрыгып", бетине тунук чык жыбырап ыйлаактанат, ошондо жегени аш болуп мал семирет, үргүлөп термелген чөптөр боюн жазып өсөт. (Ч.Айтматов, Тщнкщ сугат, аъгеме, 1955).

$$K_{синт} = 1 - \frac{П}{С} = 1 - \frac{3}{103} \approx 1 - 0{,}0291 \approx 0{,}9709$$

Мында 3 сүйлөмдөгү сөздөрдүн саны 103 жана синтаксистик ар түрдүүлүктүн катышы жогору, ал болжол менен 0,9709 барабар.

Адабияттар:

1. Апресян Ю.Д. Идеи и методы современной лингвистики /краткий очерк/. - М.: Просвещение, 1966.
2. Гладкий А.В., Мельчук И.А. Элементы математической лингвистики. -М.: Наука, 1969.
3. Ахманова О.С. Словарь лингвистических терминов. –М.: Сов. энциклопедия, 1969.
4. Адмони В.Г. Еще раз об изучении количественной стороны грамматических явлений //Вопросы языкознания, № 1, 1970.
5. Головин Б.Н. Язык и статистика. -М.: Просвещение, 1971.
6. Статистика речи и автома-тический анализ текста. -Л.: Наука, 1972.
7. Богданов В.В. Статистические концепции языка в речи /В кн.: Статистика речи и автоматический анализ текста. - Л., 1972.
8. Математическая лингвистика. Журн. Академии Наук СССР, Всесоюз-ный институт научно- технической информации. - М.: Наука, 1973.
9. Статистика речи и автомати-ческий анализ текста. -Л.: Наука, 1974.
10. Пиотровский Р.Г., Бектаев К.Б., Пиотровская А.А. Математическая лингвистика. -М.: Высшая школа, 1977.
11. Психосемантика слова и лингвостатистика текста. /Сост. А.П.Варфоломеев. Калининградсий государственный университет, 2000.
12. Крапивник Л.А. Методы лингвистического анализа. Учебно-методический материал. -Хабаровск, 2007.