КЫРГЫЗ РЕСПУБЛИКАСЫНЫН УЛУТТУК ИЛИМДЕР АКАДЕМИЯСЫ

АВТОМАТИКА ЖАНА МААЛЫМАТ ТЕХНОЛОГИЯЛАР ИНСТИТУТУ

Б. Н. Ельцина атындагы

КЫРГЫЗ-РОССИЯ СЛАВЯН УНИВЕРСИТЕТИ

Диссертациялык кеңеш Д 05.18.579

Кол жазма укугунда

УДК 519.688.004.912(575.2) (043.3)

КӨЧКӨНБАЕВА БУАЖАР ОСМОНАЛИЕВНА

КЫРГЫЗ ТИЛИ ҮЧҮН МОРФОЛОГИЯЛЫК АНАЛИЗАТОРДУН МОДЕЛДЕРИН ЖАНА АЛГОРИТМДЕРИН ИШТЕП ЧЫГУУ

05.13.18- математикалык моделдөө, сандык ыкмалар жана программалар комплекси

Техника илимдеринин кандидаты окумуштуулук даражасын изденип алуу үчүн жазылган диссертациянын

АВТОРЕФЕРАТЫ

Бишкек -2019

Диссертациялык иш академик М.М.Адышов атындагы Ош технологиялык университетинин "Эсептөө техникаларын жана автоматташытырган системаларды программалык жабдуу" кафедрасында жана А.С. Джаманбаев атындагы жаратылыш байлыктары институтунда аткарылды.

Илимий жетекчи: физика-математика илимдеринин доктору, профессор
Сатыбаев Абдуганы Джунусович
(М.М. Адышев атындагы ОшТУ, «Маалымат технологиялары жана башкаруу» кафедрасынын башчысы)

Расмий оппоненттер: физика-математика илимдеринин доктору,
КР УИАнын мүчө-корреспонденти
Панков Павел Сергеевич
(КР УИАнын Математика институту, "Эсептөө математикасы" лабораториясынын башчысы)

техника илимдеринин доктору, профессор
Торобеков Бекжан
(И. Раззаков атындагы КМТУ,
өнүгүү жана мамлекеттик тил боюнча проректор)

Жетектөөчү мекеме: **Ош мамлекеттик университети,**
"Программалоо" кафедрасы
Ош шаары, 723500, Ленин көчөсү 331

Диссертация 2019-ж. 28-июнунда саат 14:00дө Кыргыз Республикасынын Улуттук илимдер академиясынын Автоматика жана маалыматтык технологиялар институтунун жана Б.Н. Ельцин атындагы Кыргыз-Россия Славян университетинин алдындагы Д. 05.18.579 диссертациялык кеңешинин отурумунда жакталат. Дареги: 720071, Бишкек ш., Чүй проспекти, 265, ауд. 346, сайт: www.iait.kg.

Диссертация менен Кыргыз Республикасынын Улуттук илимдер академиясынын китепканасында 720071, Бишкек ш., Чүй проспектиси, 265, «а» дареги боюнча жана КР УИА АжМТИнин сайтында www.iait.kg дареги боюнча таанышса болот. E-mail: gulsaat@mail.ru.

Автореферат 2019-жылдын 27-майында жөнөтүлдү.

Диссертациялык кеңештин
окумуштуу катчысы, ф-м.и.к. _____ Керимкулова Г.К.

ИШТИН ЖАЛПЫ МҮНӨЗДӨМӨСҮ

Маселенин актуалдуулугу. Азыркы күндө, маданияттын, илимдин жана техниканын өнүгүү доорунда, маалыматтар менен иштөө: аларды өзгөртүү жана башка формага өткөрүү амалдары адам баласынын бардык чөйрөсүндө алдыңкы маселелердин бири болууда. Негизинен табигый тилдеги маалыматтарды иштеп чыгуу усулдары жана каражаттары — эң жөнөкөй документтерден баштап, маалымат издөөчү системаларга, машиналык которууларга жана компьютер менен адамдын ортосундагы табигый тилдеги пикир алмашуучу системаларга чейин чоң мааниге ээ болууда.

Табигый тилдеги тексттер менен байланышкан программалар өтө көп жана алардын берилген текстти изилдөө тереңдиги да ар түрдүү болуп эсептелет.

Берилген текстти изилдөө тереңдигине жараша түзүлгөн программалар негизинен табигый тилдеги тексттердин үстүнөн жүргүзүлүүчү төмөнкү программалык бөлүктөрдөн турат: кийирилген тексттеги тамгалар – сөздөр – сөздүн маңызы – ... - сөздүн мааниси - түзүлүшү – чыгуучу тексттин тамгалары ж.б. Кандай гана текст болбосун, аны изилдөөдө жана алгоритмин түзүүдө эң биринчи модуль катары морфологиялык анализатордун каралышы талашсыз.

Табигый тилдеги тексттердин үстүнөн иштөөчү программалардын дээрлик көпчүлүгүндө морфологиялык анализ бөлүгү эң керектүү бөлүк болуп саналат жана бул маселенин актуалдуулугун туюнтат. Программанын морфологиялык анализдөө бөлүгү изилдөөчү тексттин көлөмүнүн чоң болгонуна карабай эффективдүү жана тез иштөөсү система тарабынан талап кылынат.

Түзүлүүчү алгоритмдерге, белгилүү алгоритмдерге караганда бир канча каттуу талаптар коюулат жана азыркы күндөгү эсептөө техникаларында эстин аз бөлүгүн пайдалануусу, жогорку ылдамдыкта иштөөсү талап кылынат.

Бул багытта, кыргыз тилинин морфологиялык анализаторунун алгоритмдери кеңири изилдене элек, ошондуктан кыргыз тилинин формалдык грамматикасын, машиналык которуу, морфологиялык, синтаксистик жана семантикалык анализаторлордун негизинде эксперттик системалар сыяктуу колдонмо программаларды иштеп чыгуу **актуалдуу маселе** боюнча калууда.

Диссертациялык иштин илимий программалар жана илимий изилдөөчү иштер менен байланышы.

Диссертациялык иш академик М.М. Адышев атындагы Ош технологиялык университетинде эл аралык TEMPUS -544319-TEMPUS-1-2013-1-FR-TEMPUS-JPCR "Professional Master's Degree in computer science as a second competence in Central Asia" (PROMIS) долбоорунун алкагында ишке ашты.

Изилдөөнүн максаты. Бул диссертациялык иштин максаты принциптерди, алгоритмдерди, программаларды иштеп чыгуу жана персоналдык компьютердин оперативдик эсине кайрылууну оптималдаштыруучу лингвистикалык моделдерди камтыган кыргыз тилинин автоматтык морфологиялык анализинин математикалык моделин,

алгоритмдерин түзүү жана түзүлгөн модулду керектүү текст менен иштөөчү колдонмо программаларда пайдалануу болуп саналат.

Изилдөөнүн маселелери:

- Заманбап эсептөөчү машиналарда эффективдүү иштеп жаткан морфологиялык анализатор программаларына обзор берүү;
- Кыргыз тилинин морфологиялык таблицасын жана сөздүгүн түзүү;
- Морфологиялык анализдин маселелерин чечүүдө маалыматтар базасынын структурасын иштеп чыгуу;
- Кыргыз тили үчүн морфологиялык анализатордун математикалык моделин түзүү;
- Иштелип чыккан математикалык моделдин негизинде так жана толук морфологиялык анализ жүргүзүүчү программанын алгоритмин түзүү. Ошондой эле моделдин негизинде сөздөрдү нормалдаштыруу маселелерин изилдөө;
- Embarcadero RAD Studio XE3 чөйрөсүндө морфологиялык анализатордун программасын түзүү.

Иштин илимий жаңылыгы:

- Эрежелердин жардамында сөз формаларын белгилүү сандагы бөлүктөргө бөлө турган кыргыз тилинин морфологиялык түзүлүшүнүн модели иштелип чыкты. Бул моделдин жардамында сөздөрдү нормалдаштыруунун оптималдуу алгоритми түзүлдү;
- Биринчилерден болуп сөздүктөгү сөздөрдүн негизин издөө алгоритминин салыштырмалуу жогорку ылдамдыкта иштөөсүнө жетүүгө боло турган сөздүктүн структурасы жана мындай сөздүктөн сөздөрдү издөөнүн алгоритми иштелип чыккан. Жогорку ылдамдыкка жетүү үчүн убактылуу массив түзүлүп алынды;
- Морфологиялык анализатордун математикалык модели жана анын негизинде "NLP" программасынын иштөө алгоритми түзүлгөн. Алгоритмге ылайык атайын лексикасы жана жалпы колдонуучу 15 миңге жакын лексемди өзүнө камтыган кыргыз тилинин машиналык морфологиялык сөздүгү түзүлгөн.

Иштин практикалык мааниси — жүргүзүлгөн изилдөөлөрдүн натыйжасында персоналдык компьютердеги сөздөрдү нормалдаштыруучу автоматтык морфологиялык анализди ишке ашыруучу процедуралардын модулу пайда болду. Бул программалык каражат персоналдык компьютердеги ишке ашырылган баарлашуучу, издөөчү жана иш кагаздарды даярдоочу системаларынын эффективдүүлүгүн олуттуу жогорулатууга жол берет, жана келечекте тексттерди лингвистикалык иштеп чыгууга ар кандай статистикаларды чогултуу, издөө жана ар түрдүү шарттар боюнча тексттен фрагменттерди бөлүү, машиналык которуу системаларында жана башка ушул сыяктуу системаларга каражат катары пайдаланылат.

Иштелип чыккан жумуштун жыйынтыктары Ош гуманитардык педагогикалык институтунун электрондук библиотекасынын маалымат издөө программаларына, Ош технологиялык университетинин окуу процессине киришгелди, ошондой эле Кыргыз Республикасынын Президентине караштуу

Мамлекеттик тил боюнча улуттук комиссиясынын эксперттери тарабынан текшерилди жана жакшы баага арзыды.

Алынган жыйынтыктардын экономикалык эффективдүүлүгү.

Диссертациялык эмгектин жыйынтыгында кыргыз тили үчүн морфологиялык анализатордун программасы түзүлдү. Анын экономикалык эффективдүүлүгү эсептөөлөргө таянып 35465,52 сомду түздү.

Коргоого коюлуучу негизги жоболор:

1. Морфологиялык анализатордун иштөөсүнүн концептуалдык схемасы;
2. Морфологиялык анализди оптималдаштыруунун алгоритми;
3. Атоочтуктардын грамматикалык формаларынын фрейм-модели;
4. Кыргыз тилинин морфологиялык анализинин математикалык модели;
5. Морфологиялык анализатордун иштөө алгоритми.

Изденүүчүнүн жеке салымы:

Морфологиялык анализатордун математикалык модели жана программанын алгоритми, ошондой эле диссертациялык эмгекте келтирилген илимий жаңылыктарга ээ болгон негизги жыйынтыктар жеке автор тарабынан алынды. Кыргыз тилинин морфологиялык анализин изилдөө боюнча илимий кеңештер фил. и.д., профессор Т. Садыков тарабынан берилди жана диссертациялык иштеги маселенин коюулушу, алынган жыйынтыктарга баа берүү илимий жетекчи Сатыбаев А. Дж. менен чогуу ишке ашты.

Изилдөөнүн натыйжаларынын тастыкталышы: Автор тарабынан иштин негизги илимий жыйынтыктары "Turklang 2016" (Бишкек, 2016) IV Эл аралык конференциясында, "Turklang 2017" (Казань,2017) V Эл аралык конференциясында, М.Тагаевдин 75-жылдыгына арналган конференцияда (Бишкек,2017), А.Асановдун 75-жылдыгына арналган конференцияда (Ош,2018), "Turklang 2018" (Ташкент, 2018) VI Түрк тилдерин компьютердик иштеп чыгуу боюнча эл аралык конференциясынын алкагында доклад жасалган.

Публикациядагы диссертациянын жыйынтыктарынын толук чагылдырылышы. Диссертациялык иштин материалдары боюнча 13 макала жарык көрдү. Алардын ичинен 5 макала РИНЦке кирген Россиялык журналдарда, 6 макала Кыргыз Республикасынын ЖАК тарабынан сунушталган журналдарда басылып чыккан. Автор тарабынан иштелип чыккан "NLP Морфологический анализатор" программасына Кыргызпатен тарабынан автордук күбөлүк берилген.

Диссертациянын көлөмү жана түзүлүшү: Диссертация киришүүдөн, беш баптан, жалпы жыйынтыктан, 93 аталышты камтыган пайдаланылган адабияттардын тизмесинен жана 4 тиркемеден турат. Эмгектин көлөмү 137 беттен туруп, 9 таблицаны жана 33 сүрөттү камтыйт.

Автор диссертациялык иште жыйынтыктарды алууда баалуу кеңешин берген жетекчиси физика-математика илимдеринин доктору, профессор Абдуганы Джунусович Сатыбаевке терең ыраазычылык билдирет.

ИШТИН НЕГИЗГИ МАЗМУНУ

Киришүүдө диссертациялык иштин темасынын актуалдуугу такталды, иштин максаты, милдеттери, илимий жаңылыгы, изилдөөнүн жыйынтктарынын практикалык маанiси, ошондой эле коргоого коюулуучу негизги жоболор келтирилди.

Биринчи бапта табигый тилдеги тексттерди морфологиялык иштеп чыгуунун ыкмаларын жогорку деңгээлге жеткирүүнүн керектүүлүгү каралат. Көйгөйдү чечүүгө бар болгон жакындоолор, колдонулган усулдар жана учурда иштелип чыккан морфологиялык системалар каралды, алардын кемчиликтери менен артыкчылыктары анализденди. Изилдөөнүн милдеттери такталды.

Бул багытта кыргыз окумуштууларынын арасынан Т. Садыковдун, Н. Исраиловanын жана П.С. Панковдун эмгектерин айтууга болот.

Казак окумуштуулары А. А. Шарипбаев, Д.Р. Рахимова, У.А. Тукеев, Ж.М. Жумановдор казак тилинин мисалында орус тилинен казак тилине которуучу программанын үстүнөн иштешкен. Мында алар которууда атайын сөздүктөр таблицасын түзүп алышкан жана окшоштуктарды анализдешкен.

Орус окумуштууларынын ичинен бул маселеге кайрылгандар Г.Г. Белоногов, Е.И. Анно, О.Б. Бабко-Малая, И.А. Батманов, И.Г. Бидер, И.А. Большаков, Д. Варга, В.Н. Волков, А.Ф. Гельбух, Е.Р. Добрушина, Х.Ф. Исхакова, Е.А. Казаков жана В.А. Тузовдорду айтууга болот. В.А. Тузов эрежелерге таянган орус тилинин формалдык моделин түзүп чыккан. Алардын эмгектеринде флективдүү тилдердин анализи каралган.

Ал эми түрк тилдеринин ичинен кеңири изилдөөлөр K. Altintaş жана İ. Çiceklire, татар тилин изилдөөгө Д. Сулаймановдун жана А. Гатиатулиндин эмгектери, туркмөн тилин изилдөөгө А.С. Тантугдун жана башкалардын эмгеги таандык.

Окумуштуулар J. Hankamer, L. Karttunen, Koskenniemi, H. Trost немец жана англис тилдерин изилдешкен.

Экинчи бапта коюулган маселени чечүүдө колдонулуучу ыкмалар жана материалдар келтирилди.

Изилдөө объектиси: табигый тилдеги тексттерди морфологиялык иштеп чыгуунун методдору изилдөөнүн объектиси болуп эсептелет.

Изилдөөнүн предмети: текстти иштеп чыгуунун автоматтык морфологиялык анализинин программасын түзүүдө, аны формалдык түрдө кароо үчүн керек болгон кыргыз тилинин морфологиялык түзүлүшүн изилдөө; сөздүктү сунуштоонун ыкмалары жана компьютердин эсинде сакталган сөздүккө жеткиликтүүлүктү ылдамдатууга байланыштуу морфологиялык маалыматтар; морфологиялык анализдин алгоритмдери жана сөздөрдү нормалдаштыруу.

Изилдөө ыкмалары: алдыда коюлган маселени чечүүдө морфологиялык анализдин ыкмалары, лингвистикалык мыйзам ченемдүүлүктөрдү туюндуруучу математикалык моделдөө элементтери, ошондой эле объектке багытталган программалоонун ыкмалары колдонулган.

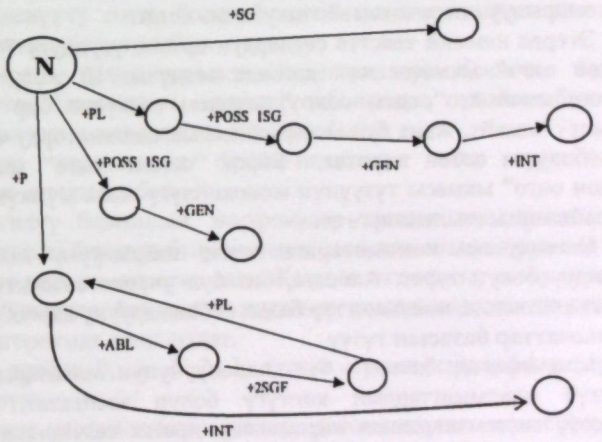
Үчүнчү бапта автор тарабынан иштелип чыккан компьютердин эсине кайрылууну азайтуу менен сөздү анализдөө үчүн баардык маалыматтарды

алууга жол берген морфологиялык сөздүктү уюштуруу ыкмасы айтылат. Сөздүктүн түзүлүшүнүн алгоритмдери менен андагы маалыматтарды издөө келтирилет.

Морфологиялык анализаторду түзүүнүн эң кеми үч жолу бар экени белгилүү: (1) сөздүккө негизделген анализаторду түзүү, (2) сөздүктү пайдаланбай, грамматикага негизделген анализаторду түзүү, жана (3) сөздүк менен грамматиканы айкаштыра колдонгон анализаторду түзүү.

Белгилүү болгондой, кыргыз тили мүчөлөмө тилдердин катары кирип, сөздүктөгү жаңы сөз кураңды мүчөнүн сөздүн негизине жалгаша келиши аркылуу жасалса, тексттеги сөздүн грамматикалык формасы сөздүктөгү сөзгө уланды мүчөнүн уланышы аркылуу уюшулат. Мында сөздүн үнгусу үндөштүктүн эрежесине ылайык өзүнө жалганган мүчөлөрдүн ар түрдүү вариантта өзгөрүп келишин шарттайт. Мисалы: тоо+Ø=тоо, тоо+нын=тоонун, үй+нын= үйдүн ж.б.

Кыргыз тилинде атооч сөз формалары сөзгө мүчөлөрдүн төмөнкү фрейм-модели боюнча уланышы аркылуу ишке ашырылат (1-сүрөт). Көрүнүп тургандай, моделдин түйүндөрү атооч сөздөрдүн морфологиясына тиешелүү ар кандай абалдары көрсөтсө, түйүндөрү байланыштырып турган өтмөктөр (жебечелер) атооч сөздөрдүн морфологиясына тиешелүү конкреттүү категорияларды көрсөтөт.Ал эми N – атооч сөздүн (noun) сөздүктөгү формасы (негизи).



1-сүрөт. Атооч сөз формаларын уюштуруунун фрейм-модели

Жогорудагы моделдин негизинде *word(x)* предикатын карай турган болсок, ал х объекти менен мүчөлөр көптүгүнүн ортосундагы байланышты камсыз кылат. Мисалдагы учурду карасак төмөнкүдөй предикаттар келип чыгат:

word(N,SG); word(кumen, SG)
word(N,P, POSS_1SG,GEN,INT);word(кumen,P, POSS_1SG,GEN,INT);

word(N, POSS_1SG,1GEN);word(кимен, POSS_1SG,1GEN);
word(N, PL, ABL, 2SGF, PL, INT); word(кимен, PL, ABL, 2SGF, PL, INT)

Ошентип, табигый тилдеги текст үчүн морфологиялык анализатордун иштөө принциптерин кароодо төмөнкүдөй этаптарды эске алуу кажет:

1. Киргизилген текстти сөз формаларына бөлүп алуу.

2. Андан соң сөз формаларын лемматизациялоо, атап айтканда, сөз формасын сөздүн сөздүктөгү формасына айландыруу.

3. Сөз формасын уюштурган уланды мүчөнү же мүчөлөрдүн тизмегин бөлүп алуу.

4. Уланды мүчөлөр тизмегин андан ары жиктөө жана ар бир мүчөнүн тиешелүү морфологиялык белгилерин аныктоо.

Маселени чечүүнүн эки жолу бар: биринчиси кийирилген сөздү «оңдон солго» карай анализдөө, экинчиси «солдон оңго» карай анализдөө ыкмасы. Айтылган ыкмалардын алгоритмдерин карап көрсөк, төмөнкүдөй жыйынтыкка ээ болобуз. Биринчи ыкмада аффикстердин комплексине окшогон акыркы бөлүктү бөлуп алуу аракети көрүлөт, андан кийин сөздүктөн калган баштапкы бөлүктү текшерүү жүргүзүлөт. Экинчи ыкмада сөздүктөн мүмкүн болгон баштапкы чынжырчаны издөө жүргүзүлөт, андан кийин калган оң жактагы бөлүк аффикстер бөлүгү катары каралат. Эки учурда тең издөө ийгиликсиз болсо, сөз формасын башынан бөлүктөө жүргүзүлөт.

Эреже сыяктуу, морфологиялык анализде сөздүктөн маалыматты аныктоо бир канча көп убакытты талап кылат. Тактап айтканда, бул учурда дисктин эсине кайрылуу операциясына көп убакыт кетет.

Эгерде илимий текстте сөздөрдүн орточо узундугу 7-10 тамгадан турса, "солдон оңго" ыкмасы жок дегенде сөздүккө 10 жолу кайрылуу жасайт. Эрежеге ылайык, "оңдон солго" ыкмасы сөздүккө бир кыйла аз сандагы кайрылуу жасайт, жана бул морфологиялык анализаторду түзүүдө тандалуучу ыкма болууга себеп жаратат. Бирок "оңдон солго" ыкмасына караганда, "солдон оңго" ыкмасы түзүүнүн жонокойлүгү жана мүмкүнчүлүктөрү боюнча бир кыйла артыкчылыктарга ээ.

Ошондуктан компьютердин эсине кайрылууну азайтуу маселеси да актуалдуу болуп турат. Албетте, биз бул учурда сөздүктү түзүү маселесин, башкача айтканда, маалыматтар базасын башкарууну каробуз туура болот.

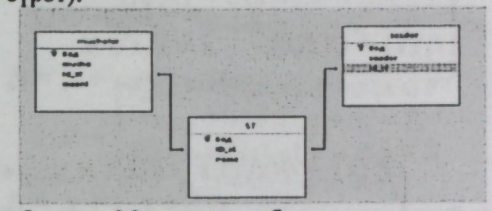
Маалыматтар базасын түзүү

Маалыматтар базасы – бул колдонуучунун талаптарын канааттандырган керектүү маалыматтардын көптүгү болуп эсептелет. Бул маалыматтар башкаруу системаларынын жардамында иретке келтирилип, таблица түрүндө сакталат жана ар бир таблица ар башка маалыматты сактайт.

Бүгүнкү күндө маалыматтарды башкаруу үчүн көптөгөн системалар колдонулат. Анткени көп сандаган маалыматтардын ичинен керектүүсүн бөлуп алуу бир канча убакытты талап кылат. Ошондуктан азыр SQL, MySQL, Oracle, Access жана башка көптөгөн маалыматтар базасын башкаруу системалары колдонулуп келет. Бул системалар аркылуу маалыматты башкаруу, сорттоо, издөө жеңил жана көп убакытты талап кылбайт. Биздин учурда табигый тилдеги текст менен иштөөчү жогорудагы алгоритмди текшерүү максатында

Embarcadero RAD Studio чөйрөсүндө тесттик программа түзүлдү. Мында маалыматтар базасын башкаруу үчүн Access системасы пайдаланылды.

Мында үч таблица түзүлүп, алардын ортосундагы байланыш төмөнкүдөй турдө түзүлдү (2-сүрөт):



2-сүрөт. Маалыматтар базасынын схемасы

Жогорудагы схемадан көрүнүп тургандай, маалыматтар базасы мүчөлөр, сөздөр жана сөз түркүмүн камтыган таблицалардан турат. Алар it_st ачкыч талаанын жардамында байланышат.

Морфологиялык анализдин алгоритмин оптималдаштыруу

Компьютердин эсине кайрылууну азайтуу максатында биз төмөнкүдөй алгоритмди пайдаландык:

1. Кийирилүүчү же анализденүүчү сөздүн биринчи эки тамгасы боюнча маалыматтар базасынан сөздөрдү бөлуп алабыз.
2. Виртуалдык эсти пайдаланып атайын массивге жайгаштырабыз.
3. Анализденүүчү сөздү "оңдон солго" ыкмасын пайдаланып салыштырып, сөздүн негизин бөлуп алабыз.
4. Мүчөлөрдү грамматикалык категориялар боюнча белгилөө үчүн маалыматтар базасындагы мүчөлөр таблицасын пайдаланабыз. Итерациянын саны мүчөлөрдүн санынан көз каранды болот.
5. 1-сүрөттө көрсөтүлгөн фрейм-моделдин текст түрүндөгү формасын жыйынтык катары алабыз

Белгилүү болгондой, морфемалар тилдин эң кичине маани берүүчү (семантикалык) бирдиги болуп саналат, алардан сөздүн формасы түзүлөт, андан ары, ошого жараша, лексема дагы. Кыргыз тилинде мүчөлөр төрт негизги түргө бөлүнөт. Төмөндө баяндалган мүчөлөр сөздүн негизин аныктоочу иштелип жаткан алгоритмде колдонулат.

Белгилеп алабыз P_i, $i= 1, 2, 3, 4$ үчүн мүчөлөрдүн (аффикстердин) төмөнкү көптүктөрүн:

Терминалдар массалык баш мүчөлөрдүн төмөнкү топтомун билдирет

P_1 – үч тамгалуу мүчөлөрдүн көптүгү (көптүк түрдүн мүчөсү);

P_2 – мүчөлөрдүн көптүгү (илик жөндөмөсүнүн мүчөлөрү);

P_3 – мүчөлөрдүн көптүгү (жеке жалгоолору);

P_4 – мүчөлөрдүн көптүгү (жөндөмө мүчөсү).

Эгерде сөз $x \in P_i$ болсо, анда бардык i = 1, …, 4 үчүн $P_i(x)$ деп алабыз.

Эгерде сөз $x \in W$ болсо, анда $W(x)$ деп белгилейбиз.

Эгерде сөз $x \in Q$ болсо, анда $Q(x)$деп белгилейбиз.

Анда биздин A–H эрежелеribиз төмөнкү формулалар боюнча өзгөрөт [6].

Каалагандай z сөзү $x_0 + x_1 + x_2 + \ldots + x_k$ тамгалардан турсун дейли. (x_i –сөздөгү тамгалардын эң көбү). $i = k$, $x = x_i$ десек:

1-кадам.

$$A = \begin{cases} P_4(x) \to Q(z\backslash x), & z = z\backslash x \in N, x_{i-n} + \cdots + x_{i-1} + x_i \neq P_1 \\ & n = length(P_1) \\ P_1(x) \to Q(z\backslash x), & z = z\backslash x \in N, \\ & x_{i-2} + x_{i-1} + x_i \neq P_4 \end{cases}$$

2-кадам

$$B = \begin{cases} P_2(x) \to Q(z\backslash x), z = z\backslash x \in N, x_{i-2} + x_{i-1} + x_i \neq P_1 \\ \qquad x_{i-2} + x_{i-1} + x_i \neq P_4 \\ P_1(x) \to Q(z\backslash x), z = z\backslash x \in N, \\ \qquad x_{i-2} + x_{i-1} + x_i \neq P_4 \end{cases}$$

3-кадам

$$C = \begin{cases} P_4(x) \to Q(z\backslash x), z = z\backslash x \in N, \\ \qquad x_{i-2} + x_{i-1} + x_i \neq P_1 \\ P_3(x) \to Q(z\backslash x), z = z\backslash x \in V \\ P_1(x) \to Q(z\backslash x), z = z\backslash x \in N, \\ \qquad x_{i-n} + \cdots + x_{i-1} + x_i \neq P_3 \\ \qquad n = length(P_3) \end{cases}$$

4-кадам

$$D = \begin{cases} P_2(x) \to Q(z\backslash x), z = z\backslash x \in N, \\ \quad x_{i-n} + \cdots + x_{i-1} + x_i \neq P_1 \\ P_3(x) \to Q(z\backslash x), z = z\backslash x \in V, \\ \quad x_{i-n} + \cdots + x_{i-1} + x_i = P_4 \\ P_1(x) \to Q(z\backslash x), z = z\backslash x \in N, \\ \quad x_{i-n} + \cdots + x_{i-1} + x_i \neq P_3 \end{cases}$$

5-кадам

$$E = \begin{cases} P_1(x) \to Q(z\backslash x), z = z\backslash x \in N, \\ \qquad x_{i-2} + x_{i-1} + x_i \neq P_3 \\ P_2(x) \to Q(z\backslash x), z = z\backslash x \in V, x_{i-n} + \cdots + x_{i-1} + x_i = P_1 \\ \qquad x_{i-n} + \cdots + x_{i-1} + x_i \neq P_4 \end{cases}$$

6-кадам

$$F = \begin{cases} P_4(x) \to Q(z\backslash x), & z = z\backslash x \in N, \\ \quad x_{i-n} + \cdots + x_{i-1} + x_i \neq P_1 \\ P_3(x) \to Q(z\backslash x), z = z\backslash x \in V, \\ \quad x_{i-n} + \cdots + x_{i-1} + x_i = P_4 \end{cases}$$

7-кадам

$$G = \begin{cases} P_2(x) \to Q(z\backslash x), z = z\backslash x \in V, \\ \quad x_{i-n} + \cdots + x_{i-1} + x_i = P_1 \\ P_3(x) \to Q(z\backslash x), z = z\backslash x \in V \\ \quad x_{i-n} + \cdots + x_{i-1} + x_i = P_4 \end{cases}$$

8-кадам

$$H = \begin{cases} P_3(x) \to Q(z\backslash x), z = z\backslash x \in V \\ \quad x_{i-n} + \cdots + x_{i-1} + x_i = P_4 \\ P_1(x) \to Q(z\backslash x), z = z\backslash x \in N, \\ \quad x_{i-2} + x_{i-1} + x_i \neq P_3 \end{cases}$$

Бул кадамдардан кийин сөздөрдүн аффикстерин текшерүү бүтөт, эгерде жыйынтык жок болсо 1-кадамга кайрадан кайтат. Жыйынтыгында сөздүн нормалдуу формасын алабыз.

Төртүнчү бап иштин борбордук бөлүгү болуп эсептелет. Мында автор тарабынан иштелип чыккан кандайдыр бир деңгээлде тилден көз каранды болбогон табигый тилдин формалдык лингвистикалык модели жана бул моделдин негизинде түзүлгөн системанын структуралык схемасы келтирилди (3-сүрөт).

Морфологиялык категориялар. Алардын тизмеси төмөнкү категориялар менен аныкталат:

1. Зат атооч – Noun:

Сан категориясы – Number

1. Жекелик сан – singular
2. Көптүк сан – plural

Энтектери:

1. **SG <=>Ø**
2. **PL <=> ЛАр**

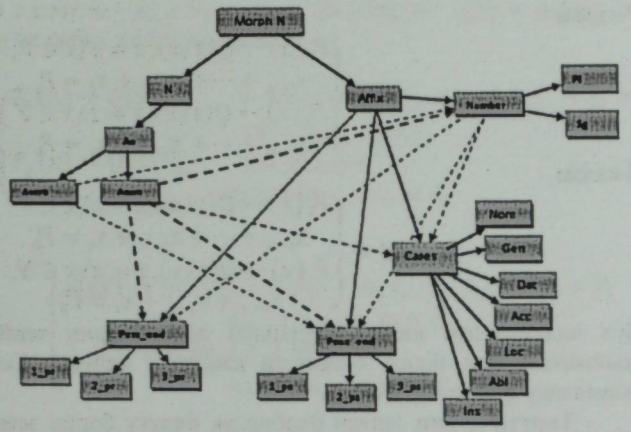
Таандык категория – Possessive

Жекелик сан – singular:

1. Таандык жекелик сан 1-жак- 1st person singular possessive ('my'),
2. Таандык жекелик сан 2-жак- 2nd personsingularpossessive ('your'),
3. Таандык жекелик сан 2-жак, сылык түрү- 2nd person sing.poss. formal ('your'),
4. Таандык жекелик сан 3-жак- 3nd person singular possessive ('his/her/its'),

Көптүк сан – plural:

5. Таандык көптүк сан 1-жак - 1st person plural possessive ('our'),
6. Таандык көптүк сан 2-жак - 2nd person plural possessive ('your'),
7. Таандык көптүк сан 2-жак, сылык түрү - 2nd person pl.poss. formal ('your'),
8. Таандык көптүк сан 3-жак - 3nd person plural possessive ('their'),



3-сүрөт. Атоочтук сөздөрдүн жасалуу модели

Кыргыз тилинин морфологиясынын математикалык модели

Ар кандай агглютинативдик тилдер үчүн сөз формасын $S_n = h_1 h_2 \dots h_n$ деп алсак, мында h_i *(i=1,2,...,n)* A алфавитиндеги элемент, n – элементтердин саны (жолчонун узундугу). Биз изилдөөдө кыргыз алфавитин пайдаланабыз. Алфавит 36 тамгадан турат жана "_" символу курук символ үчүн колдонулат:

A={а,б,в,г,д,е,ё,ж,з,и,й,к,л,м,н,ң,о,ө,п,р,с,т,у,ү,ф,х,ц,ч,ш,щ,ъ,ы,ь,э,ю,я,э, '_'}

жана биз жолчого камтылуучу жолчо үчүн төмөнкүдөй белгилөөнү кийирип алабыз: $1 \le i \le j \le n$ үчүн S_n камтылуучу жолчосу төмөнкүдөй аныкталат:

$$S_n[i:j] = h_i h_{i+1} \dots h_j$$

$$S_n[:j] = h_1 h_2 \dots h_j$$

$$S_n[i:] = h_i h_{i+1} \dots h_n$$

Биздин белгилөөлөрдүн негизинде атайын камтылуучу жолчо $S_n[i:i+1] = h_i h_{i+1}$ удаалаш тамгалар жубу $(h_1,h_2)_i$ аркылуу белгиленет. Мында камтылган индекс i *(i=1,2,...,n-1) ге* барабар жана удаалаш жайгашкан тамгалар жубунун баштапкы позициясын көрсөтөт $h_1=h_i$, $h_2=h_{i+1} \in A$. $i=n$ үчүн удаалаш жуп null символоду кошуу менен $(h_n, '\_')_{i=n}$ барабар болот. Ошентип $S_n = h_1 h_2 \dots h_n$ жолчосу n удаалаш жупка ээ болот.

$1 \le j \le n_{max}$ *(мында* n_{max} кыргыз тилиндеги мүмкүн болгон сөз узундугу*)* интервалында жолугуусу мүмкүн болгон берилген иреттелген $(h_1,h_2)_j$ тамгалар жубу үчүн жана $S_n = h_1 h_2 \dots h_n$; $n \ge j$болсо, $(h_1,h_2)_j \in S_n$ учурда удаалаш жуп болгон $(h_1,h_2)_i$ S_nдин i $(1 \le i \le n)$ позициясында болсо, $(h_1,h_2)_i = (h_1,h_2)_j$ $i=j$ үчүн аткарылат. Жыйынтыктап айтканда биз дагы эки символду таптык:

$g_m=S_n[:m]$ *жана* $e_m=S_n[m:]$ ар кандай сөз формасын иреттелген эки жуп камтылуучу жолчолор катары төмөнкүдөй берүүгө болот:

$$\text{бардык } 1 \le m \le n \text{ үчүн } S_n^m = (g_m, e_m).$$

L көптүгү $(h_1,h_2)_i$ иреттелген жуп тамгаларынын мүмкүн болгон бардык түрүн камтысын жана каалаган кыргыз сөздөрүндө $i=1,\dots n_{max}$ позициясында жолуксун. Анда L жөнөкөй мейкиндиги төмөнкүдөй аныкталат:

$$L = \{(h_1,h_2)_i | h_1,h_2 \in A \; and \; 1 \le i \le n_{max}\}$$

Андан ары G_k, E_k, жана T_k көптүктөрү берилсин. Мында G_k, E_k, $T_k \subset L, 1 \le k \le n_{max}$ абалды аныктайт жана төмөнкүдөй табылат:

$$G_k = \{(h_1,h_2)_i | i = k \; and \; (h_1,h_2)_i \in g_m \; and \; 1 \le m \le n_{max}\}$$

$$E_k = \{(h_1,h_2)_i | i = k \; and \; (h_1,h_2)_i \in e_m \; and \; 1 \le m \le n_{max}\}$$

$$T_k = \{(h_1,h_2)_i | i = k, \; h_1 = s_n[k:k], \; h_2 = s_n[k+1,k+1], \; 1 \le i \le n_{max}\}$$

Ошентип, $(h_1,h_2)_i$ иреттелген жубу үчүн $S_n = h_1 h_2 \dots h_n$ менен белгиленген ар кандай берилген сөз формасынын $i=1,2,\dots,n$ позициясында жолугуусу жогоруда каралган үч көптүктөн төмөнкүдөй аныктоого болот:

$$Pr(s_n[i:i+1] \in G_i) = Pr((h_1,h_2)_i \in G_i) = P_G((h_1,h_2)_i) \quad (1)$$

$$Pr(s_n[i:i+1] \in E_i) = Pr((h_1,h_2)_i \in E_i) = P_E((h_1,h_2)_i) \quad (2)$$

$$Pr(s_n[i:i+1] \in T_i) = Pr((h_1,h_2)_i \in T_i) = P_T((h_1,h_2)_i) \quad (3)$$

Мында (1) теңдеме иреттелген $(h_1,h_2)_i$ жубунун сөздүн негизинде жайгашуусун аныктайт, (2) теңдеме иреттелген $(h_1,h_2)_i$ жубунун сөздүн аффикс бөлүгүндө жайгашуусун аныктайт, ал эми (3) теңдеме иреттелген $(h_1,h_2)_i$ жубунун сөздүн негизи менен аффикс бөлүктөрүнүн ортосунда жайгашуусун аныктайт.

Кыргыз тилиндеги сөздөрдүн грамматикалык формасы агглютинативдик чынжырчанын жаралуу эрежесине ылайык уңгудан жана уланды аффикстерден турат. Сөздөрдүн грамматикалык формасынын моделин төмөнкүдөй көрсөтүүгө болот:

$$S = R + \sum_{i=0}^{m} U_i, (m \le 8) \quad (4),$$

мында, S– сөз формасы, R –уңгу же сөздүн негизи, U_i- уланды аффикстер.
(4) формуладан көрүнүп тургандай сөздүн грамматикалык формасын табууда *k=length(S)-length (R)* жолу итерациялоодон кийин сөздүн уңгусу жана уланды мүчөлөр бөлүнүп алынат.

Уланды аффикстердин саны сегиз көрсөткүчкө чейин жетиши ыктымал, башкача айтканда:

$$\sum_{i=0}^{8} U_i = U_0 + U_1 + U_2 + \dots + U_8 \quad (5)$$

Аныктама 1: Эгерде $U_m=\varnothing$ болсо, анда S функциясы сөздүн негизине же уңгуга барабар болот, жана кийирилген сөз эч кандай ажыроосуз сөздүктөгү лексемага дал келет.

Эгерде изилденүүчү сөз формасынын узундугу *l=length(S)* аркылуу табылса, анда сөздүктөн табылган сөз формаларынын көптүгүн m деп алсак, ал (lxm) сегментациялык матрицасын жаратат.

$$m\begin{Bmatrix} \underbrace{1\ 1\ a_{31}a_{41}\ \dots\ a_{l1}}_{i\geq l} \\ 1\ 1\ a_{32}a_{42}\ \dots\ a_{l2} \\ \dots \\ 1\ 1\ a_{3m}a_{4m}\ \dots\ a_{lm} \end{Bmatrix} \quad l$$

матрицадагы ар бир жолчо бир сөз формасын берет жана кайсы жолчонун суммасы максималдуу болсо, ошол сөздүн негизи болот.
Матрицанын негизинде

$$\begin{cases} x_1 + x_2 + x_3 + \dots + x_l \leq l \\ \dots \\ x_1 + x_2 + x_3 + \dots + x_{l-i} \leq l - i \end{cases}$$

теңдештиктер системасы алынат.
Системаны чыгарууда кайсы жолчодо бирлердин саны көп болсо, ошол жолчонун элементтеринен түзүлгөн көптүк теңдеменин тамыры катары кабыл алынат.
Мисалы, *абалы* деген сөздүн нормалдуу формасын табууда сөздүктөн 6 массивдин элементи алынат. Массивдердин элементтерин сегментациялоодон кийин компьютердин эсинде төмөнкүдөй матрица түзүлөт:

$$M=\begin{vmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 \end{vmatrix}$$

Матрицанын элементтерин анализдеп 4 жана 5 жолчолорду бөлүп алабыз:

$$M=\begin{vmatrix} 1 & 1 & 11 & 0 \\ 1 & 1 & 11 & 0 \end{vmatrix}$$

Бул эки жолчонун узундуктарын бир өлчөмдүү массивге жайгаштырабыз:
$L=\{l_1, l_2\}$. Биздин учурда эң жакын жолчолордун саны 2 ге барабар болду. Кээ бир учурда бул көрсөткүч бир кыйла көп болот. Мисалы *кыргыздар* сөзү үчүн бул көрсөткүч {кыргыз, кыргын, кыргыч, кыргый, ...} болуп өсүп кетет.
Алынган массивдин жакындаштырылган маанисин итерациялоо аркылуу табабыз:

```
For i:=1 to length(L) do
If min>l[i] then min:=l[i];
```


Бул механиздин жыйынтыгында узундугу боюнча минималдуу жолчону алабыз. Ал M[5] болуп табылат б.а. төмөнкү фрагмент аткарылат:

```
ss2:=trim(Edit1.Text);
d:=strtoint(edit3.Text);
edit2.Text:='';
1:
l:=length(ss2);
for i:=1 to 46 do
if (ss2=m[i]) or (ss2=(copy(m[i],1,length(m[i])-1))+'б')
or (ss2=(copy(m[i],1,length(m[i])-1))+'г') then
begin
edit2.Text:=m[i];
form1.RichEdit3.Lines.Add(mucho.edit2.Text);
case strtoint(m3[i]) of
1:form1.RichEdit3.Lines.Add('зат атооч');
2:form1.RichEdit3.Lines.Add('сын атооч');
3:form1.RichEdit3.Lines.Add('сан атооч');
4:form1.RichEdit3.Lines.Add('ат атооч');
5:form1.RichEdit3.Lines.Add('этиш');
6:form1.RichEdit3.Lines.Add('тактооч');
7:form1.RichEdit3.Lines.Add('сырдык сөз');
8:form1.RichEdit3.Lines.Add('туураңды сөз');
9:form1.RichEdit3.Lines.Add('жандооч');
10:form1.RichEdit3.Lines.Add('байламта');
11:form1.RichEdit3.Lines.Add('кызматчы сөз');
end;end;
ss3:=ss3+copy(ss2,l,l);
ss2:=copy(ss2,1,l-1);
if edit2.Text='' then goto 1
else if (m[i]='') then
begin
timer1.Enabled:=true; end;
aff:='';
for i:=1 to length(ss3)-1 do
aff:=aff+ss3[length(ss3)-i];edit4.Text:=aff;
```

Морфологиялык анализатордун алгоритми

Аталган алгоритмдин блок схемасы 4 жана 5-сүрөттөрдө көрсөтүлгөн.
Морфологиялык анализ жүргүзүүнүн 3 деңгээлин бөлүп карооо болот:

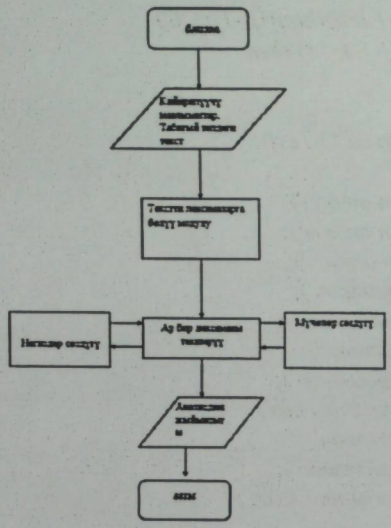
1. Сөздүн грамматикалык маанисин гана аныктоо;
2. Сөздүн негизин гана аныктоо;
3. Сөздүн грамматикалык маанисин жана негизин аныктоо.

Морфологиялык анализди өтө терең же толук эмес изилдөө коюлган маселеге жараша болот.

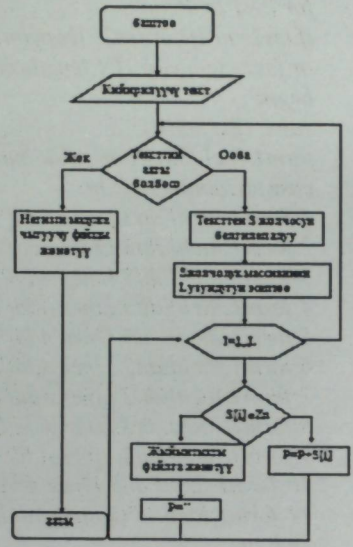
Морфологиялык анализ табигый тил менен байланышкан ар түрдүү маселелердин баштапкы тепкичи болгондуктан, анын канчалык так аткарылгандыгы чоң мааниге ээ.

Морфологиялык анализдин ыкмаларын 3 типке бөлсө болот:

1. мүчөлөрдүн сөздүгү менен анализдөө;
2. мүчөлөр жана негиздер сөздүгүнүн жардамында анализдөө;
3. сөз тутумдарынын сөздүгүнүн жардамында анализдөө.



4-сүрөт. Морфологиялык анализдин алгоритми



5-сүрөт. Текстти лексемаларга бөлүү модулунун алгоритми

Бешинчи бапта түзүлгөн морфологиялык анализатордун программасын тесттен өткөрүү амалдары каралат.

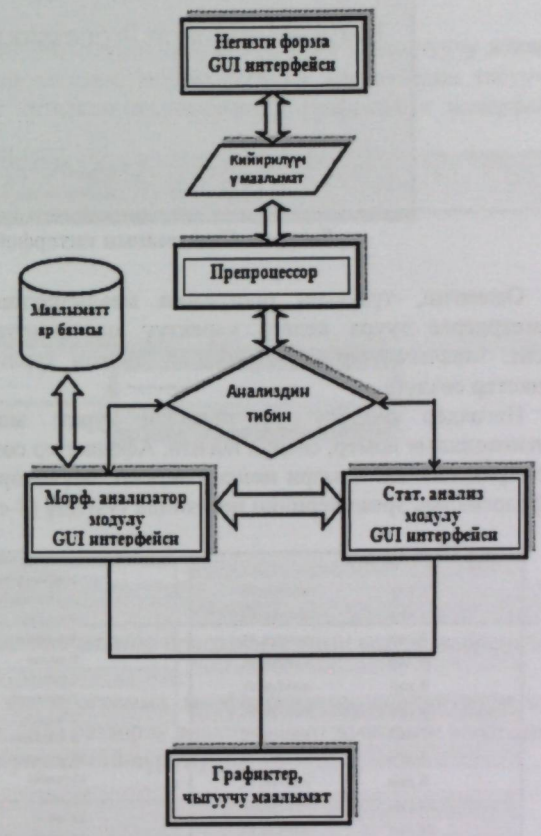
Иштелип чыккан программа табигый тилде жазылган текстерди кийирилүүчү маалымат катары кабыл алат жана анын негизин бөлүп алат. Жыйынтыгында табылган сөздүн лексикалык маанилери чыгуучу маалымат катары кайтарылат. Программаны түзүү үчүн объекттер менен иштөөчү RAD Studio XE3 чөйрөсү тандалып алынды. Бул чөйрөдө типтер менен иштөөдө кийирилген тексттерди кайра иштеп чыгуу үчүн кыргыз тилинин алфавити үчүн unistring тиби колдонулду.

Түзүлгөн система: маалыматтар базасынан, колдонуучу үчүн ыңгайлуу интерфейстен, морфологиялык анализ жана статистикалык анализ модулдарынан турат.

Программанын техникалык камсыздалуусун карасак, 800 дөн ашык жолчодон турат жана 15,7 Мб эстин көлөмүн эзлейт. Маалыматтар базасы менен иштөөдө 16 Мб эстин көлөмү керектелет жана лингвистикалык

таблицаны сактоо үчүн 40 Кб эстин көлөмү керектелет. Ошентип, "NLP" программасы 69 функциядан жана 22 туракту чоңдуктан турат.

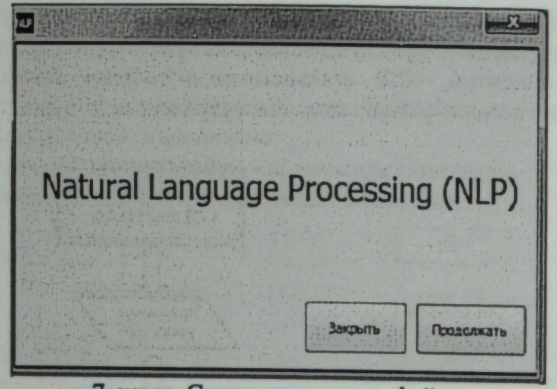
Ошентип, NLP аталышындагы система иштелип чыкты. Системаны түзүүдө концептуалдык схеманы түзүп алдык (6-сүрөт.)



6-сүрөт. Түзүлүүчү системанын концептуалдык схемасы

Түзүлгөн системаны тесттен өткөрүү

Морфологиялык анализатордун бул биринчи версиясында маалыматтар базасында көп колдонулуучу маалыматтар камтылган. Колдонуучу үчүн интерфейс маалыматтар базасынан маалыматтарды алуу максатында GUI форматында түзүлгөн (7-сүрөт).



7-сүрөт. Системанын интерфейси

Ошентип, түзүлгөн программа маалыматтар базасынан кийирилген параметрлерге туура келген керектүү маалыматтарды бөлүп алуу менен иштейт. Маалыматтар базасы эки сөздүктөн турат: негиздер сөздүгү жана аффикстер сөздүгү.

Негиздер сөздүгү үч талаадан турат: морфологиялык маалымат, идентикациялык номер, сөздүн негизи. Аффикстер сөздүгүндө мүчөлөр көптүгү идентификациялык номери менен сакталат. Мүчөлөр көптүгү кыргыз тилинин морфологиялык эрежелеринин негизинде түзүлдү (8-сүрөт).

Код	mucho	
1	лар	көптүк, PL
2	лер	көптүк, PL
3	лор	көптүк, PL
4	лөр	көптүк, PL
5	дар	көптүк, PL
6	дер	көптүк, PL
7	дер	көптүк, PL
8	дор	көптүк, PL
9	дөр	көптүк, PL
10	тар	көптүк, PL
11	тер	көптүк, PL

а)

код	sozdcr	id_st
1	аалам	1
2	аламдык	3
3	аалаш	5
4	аалим	1
5	аалы	1
6	алым	1
7	амыят	1
8	аарчы	5
9	аарчыл	5
10	аарчын	5
11	аары	1
12	аба	1
13	абаз	1
14	абай	1

б)

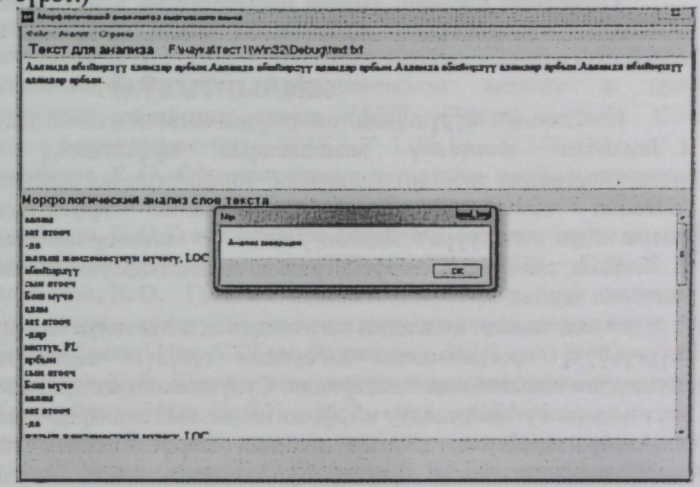
8-сүрөт. а) аффикстер базасы; б) негиздер базасы

Теориялык жактан сөз өзгөртүүчү аффикстерден түзүлгөн бул чынжырчалар агглютинативдик түрк тилдеринде чексиз узундукка ээ болуусу мүмкүн. Бирок, маалыматтар базасын түзүүдө аны чектүү деп алуу кабыл алынган жана ал эң көбү сегиз аффикске чейин жетүүсү мүмкүн жана бул статистикалык жактан негиздүү.

Морфологиялык анализатор программасынын иштөө механизми төмөнкүдөй. Морфологиялык анализдин программасы аффикстердин чынжырынын улануу мүмкүндүгүн алломорфтордун кезектешүү эрежесине

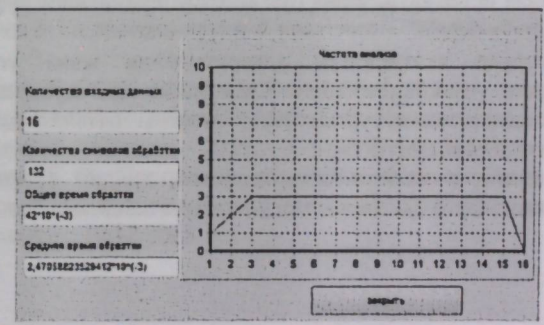
ылайык текшерет, ошондой эле алынган негиздин уланган аффикстердин негизинде морфонологиялык касиеттерге жооп берүүсү текшерилет. Программанын иштөөсү үчүн керек болгон бардык маалыматтар программанын иштөөсү менен оперативдик эске жүктөлөт. Ошентип, түйүндүк маалыматтар базасына кайрылуудан арылабыз жана программанын иштөө тездиги жогорулайт.

Кыргыз тили үчүн сүйлөмдөрдү сөздөргө ажыратуучу жана ар бир сөздү анализдеп, сөздүн негизин, морфологиялык касиеттерин табуучу программа иштелип чыкты. Алынган жыйынтыктар графикалык интерфейсте көрүнүп турат (9-сүрөт.)



9-сүрөт. Анализатор программасынын иштөө жыйынтыгы

Ошондой эле программа морфологиялык анализ үчүн статистикалык анализ да жасайт, анда сөздөрдү анализдөөнүн жыштыгы эсептелип чыгат жана диаграммасы тургузулат (10-сүрөт.)



10-сүрөт. Статистикалык анализдин жыйынтыгы

Жыйынтыктап айтканда, кыргыз тили бардык түрк тилдери сыяктуу эле агглютинативдик тилдердин катарына кирет жана сөздөр негизден жана аффикстер көптүгүнөн турат.

NLP программасын тесттен өткөрүүдө оң жыйынтыктарды берди жана ал маалымат-издөөчү программаларга, машиналык которуу системаларында баштапкы модул катары же окуу процессинде кыргыз тилинин морфологиялык анализин үйрөнүү максатында колдонулат.

Корутундуда диссертациялык иште алынган негизги илимий жана практикалык жыйынтыктар тизмектелди.

Тиркемеде иштелип чыккан программалык коддун листинги, кириштөө актылары, аффикстердин жыйындысы жана программага Кыргызпатент тарабынан алынган күбөлүк тиркелди.

ЖЫЙЫНТЫКТАР

Изилдөөнүн жүрүшүндө төмөнкүдөй илимий жыйынтыктар алынды:

1. Заманбап эсептөөчү машиналарда эффективдүү иштей турган агглютинативдик табигый тилдердин морфологиялык жактан түзүлүү модели иштелип чыкты. Түзүлгөн программанын эффективдүүлүгү катары оперативдик эсти туура пайдалануу жана тез иштөөсү эсептелет.
2. Кыргыз тили үчүн морфологиялык анализатордун математикалык модели иштелип чыкты.
3. Иштелип чыккан моделдин негизинде так жана толук морфологиялык анализ жүргүзүүчү программанын алгоритми түзүлүп чыкты жана формалдык тилдердин жардамында текшерилди. Сунушталган алгоритмдер жана моделдер жалпысынын универсалдуу морфологиялык системаларды түзө алат.
4. Автор тарабынан кыргыз тилинин морфологиялык таблицасы жана морфологиялык сөздүк түзүлдү. Бул сөздүк өз ичине (белгилүү булактардан алынган) 15 000 жакын лексеманы камтыйт.
5. Морфологиялык анализдин маселелерин чечүүдө маалыматтар базасы менен иштөө талабы келип чыккандыктан, программада маалыматтар базасынын структурасы иштелип чыкты. Маалыматтарды башкаруу үчүн компьютердин эсине бир жолу кайрылуу менен керектүү сөздөрдүн массивин түзүп алуу амалдары ишке ашты жана бул компьютердин эсин эффективдүү пайдалануу мүмкүндүк берди.
6. Жогоруда келтирилген алгоритмдерди жана усулдарды практикалык тестирлөө үчүн Embarcadero RAD Studio XE3 программалоо чөйрөсүндө 800 дөн ашык жолчодон турган "NLP" системы иштелип чыкты.

Практикалык сунуштар

Алынган жыйынтыктар табигый тилдин үстүнөн иштөөчү колдонмо программаларга баштапкы модул катары пайдаланылат.

ЖАРЫК КӨРГӨН ЭМГЕКТЕРДИН ТИЗМЕСИ

1. **Кочконбаева, Б. О.** Automatic processing of text in natural language [Текст] / Б. О. Кочконбаева, А. Алдосова // Бюллетень науки и практики. – 2018. – Т. 4, № 7. – С. 216-221.
2. **Кочконбаева, Б. О.** Алгоритм синтаксического анализатора для машинного перевода текстов [Текст] / Б. О. Кочконбаева // Труды VМеждунар. науч.-практ. Конф. Информатизация общества. – Астана, 2016. – С.92-95.
3. **Кочконбаева, Б. О.** Защита информации с помощью криптографических методов [Текст] / Б. О. Кочконбаева, Н. Р. Абдыраева // Изв. ОшТУ. – 2010. – № 2. – С.183-186.
4. **Кочконбаева, Б. О.** Лексический анализатор естественного текста [Текст]/ Б. О. Кочконбаева, Н. Р. Абдыраева // Изв. ОшТУ. – 2014. – С.207-209.
5. **Кочконбаева, Б. О.** О морфологическом анализе в приложениях автоматической обработки текста (АОТ) [Текст] / Б. О. Кочконбаева //Бюллетень науки и практики. – 2018. – Т. 1, № 12. – С.608-612.
6. **Кочконбаева, Б. О.** Об оптимизации алгоритма морфологического анализа [Текст] / Б. О. Кочконбаева, Т. Садыков. – Ташкент, 2018. – С. 293-299
7. **Кочконбаева, Б. О.**Компьютерная обработка естественного языка [Текст] / Б. О. Кочконбаева, Н. Р. Абдыраева // Изв. ОшТУ. – 2015. – С.86-89.
8. **Кочконбаева, Б. О.** Табигый тилдеги тексттерди орус тилинен кыргыз тилине машиналык которууда сөздөрдү анализдөөнүн алгоритмин түзүү [Текст] / Б. О. Кочконбаева // Изв. КТУ им. Раззакова. – 2016. – № 2(38). – С.55-58.
9. **Кочконбаева, Б. О.** Кыргыз тили үчүнсөздүннегизинаныктоо модели [Текст] / Б. О. Кочконбаева // Изв. ОшТУ. – 2018. – № 1. – С.24-30.
10. **Кочконбаева, Б. О.** Улуттук корпус үчүн морфологиялык белгилөөлөр [Текст] / Б. О. Кочконбаева, Т. С. Садыков, Б. Ш. Шаршенбиев // Вестн. КРСУ. – 2018. – Т. 18, № 1. – С.91-95.
11. **Кочконбаева, Б. О.** Модель морфологического анализа кыргызского языка [Текст] / Б. О. Кочконбаева, Т. С. Садыков // Издательство Академии наук Республики Татарстан – Казань, 2017. – С.135-155.
12. **Кочконбаева, Б. О.** Математическое моделирование и алгоритм морфологического анализа кыргызского языка [Текст] / Б. О. Кочконбаева, А. Дж. Сатыбаев // Бюллетень науки и практики. – 2019. – Т. 5, № 3. – С. 220-224.
13. **Кочконбаева, Б. О.** Тестирование программы морфологического анализатора естественного языка [Текст] / Б. О. Кочконбаева, А. Дж. Сатыбаев //Бюллетень науки и практики. – 2019. – Т. 5, № 3. – С. 215-219.
14. **Кочконбаева, Б. О.** Программа для ЭВМ «Natural Language Processing. Морфологический анализатор кыргызского языка» [Текст] / Свидетельство КР, №537-Кыргызпатент, 19.12.2018.

Кочконбаева Буажар Осмоналиевнанын 05.13.18 - математикалык моделдөө, сандык ыкмалар жана программалар комплекси адистиги боюнча «Кыргыз тили үчүн морфологиялык анализатордун моделдерин жана агоритмдерин иштеп чыгуу» аттуу темасында аткарылган диссертациясынын

ТАРЖЫМАЛЫ

Ачкыч сөздөр: морфологиялык анализатор, машиналык которуу, стемминг, морфологиялык талдоо, лемматизация, аффикстер, сөз формасы, сөздүн нормалдык формасы.

Изилдөө объектиси: табигый тилдеги тексттерди морфологиялык иштеп чыгуунун ыкмалары изилдөөнүн объектиси болуп эсептелет.

Изилдөөнүн предмети: текстти иштеп чыгуунун автоматтык морфологиялык анализинин программасын түзүүдө, аны формалдык түрдө каroo үчүн керек болгон кыргыз тилинин морфологиялык түзүлүшүн изилдөө; сөздүктү сунуштоонун ыкмалары жана компьютердин эсинде сакталган сөздүккө жеткиликтүүлүктү ылдамдатууга байланыштуу морфологиялык маалыматтар; морфологиялык анализдин жана сөдөрдү негизин табуунун ыкмалары жана алгоримдери.

Иштин максаты: морфологиялык анализатордун алгоритмдерин жана моделдерин түзүү.

Изилдөө ыкмалары: алдыда коюлган маселени чечүүдө морфологиялык анализдин ыкмалары, лингвистикалык мыйзам ченемдүүлүктөрдү туюндуруучу математикалык моделдөө элементтери, ошондой объектке багытталган программалоонун ыкмалары колдонулган.

Аппаратура: ноутбук Intel Core i3, Embarcadero RAD Studio XE3

Иштин негизги натыйжалары: морфологиялык анализдин математикалык моделдери жана алгоритмдери ишелип чыкты, ошондой эле морфологиялык анализатордун автоматташтырылган системасы жана көп колдонулуучу сөздөрдүн сөздүгү түзүлгөн.

Изилдөөнүн натыйжаларын колдонуу: иштелип чыккан морфологиялык анализатордун системасы М.М. Адышев атындагы Ош технологиялык университетинин окуу процессинде колдонууга киргизилди жана мамлекеттик Ош педагогикалык институтунун электрондук библиотекасына маалымат издөөчү модул катары кириштелди. Ошондой эле Кыргыз Республикасынын Президентине караштуу Мамлекеттик тил боюнча улуттук комиссиясынын эксперттери тарабынан жактырылды.

Колдонуу тармагы: изилдөөнүн натыйжалары жана иштелип чыккан система машиналык которуу, эксперттик системаларда, окутуу жана үйрөтүүчү системаларда базалык модуль катары колдонулат.

РЕЗЮМЕ

диссертации Кочконбаевой Буажар Осмоналиевны на тему: "Разработка моделей и алгоритмов морфологического анализатора для кыргызского языка" на соискание ученой степени кандидата технических наук по специальности 05.13.18 - математичексое моделирование, численные методы и комплексы программ

Ключевые слова: морфологический анализатор, машинный перевод, стемминг, морфологический анализ, лемматизация, аффиксы, словоформа, нормальная форма слов.

Объект исследования: методы обработки текстов естественного языка.

Предмет исследования: изучение строения словоформ кыргызского языка, создание программы автоматического морфологического анализатора для обработки естественного текста; визуализация морфологических данных, с хорошим доступом к словарю, хранящегося на жестком диске; методы и алгоритмы морфологического анализа и нормализации слов;

Цель исследования: разработка автоматизированного морфологического анализатора.

Методы исследования: при решении поставленных задач в работе использованы методы морфологического анализа. Применены элементы моделирования для построения математических моделей, описывающих лингвистические закономерности, а также методы объектно-ориентированного программирования.

Аппаратура: ноутбук Intel Core i3, Embarcadero RAD Studio XE3

Полученные результаты и их новизна: разработаны математические модели и алгоритмы морфологического анализа, а также автоматизированная система морфологического анализатора и словарь с часто используемыми словами.

Использование результатов исследования: автоматизированная система морфологического анализа внедрена в учебный процесс Ошского технологического университета им. М.М. Адышева и в электронную библиотеку Ошского государственного педагогического института в качестве модуля поиска информации. А также программа получила положительные отзывы от экспертов национальной комиссии по Государственному языку при Президенте Кыргызской Республики.

Область применения: Результаты исследования и разработанная система могут быть использованы в системах машинного перевода, экспертных системах, обучающих системах, как базисный модуль.

SUMMARY

of the dissertation of Kochkonbaeva Buazhar Osmonalievna on the theme: "Development of models and algorithms of morphological analyzer for the Kyrgyz language" for the degree of candidate of technical sciences, specialty 05.13.18 - mathematical modeling, numerical methods and program complexes.

Keywords: morphological analyzer, machine translation, stemming, morphological analysis, lemmatization, affixes, word form, normal form of words.
Object of the research: Natural language text processing methods.
Subject of research: study of the structure of word forms of the Kyrgyz language, the creation of an automatic morphological analyzer program for processing natural text; visualization of morphological data, with good access to the dictionary stored on the hard disk; methods and algorithms for morphological analysis and normalization of words;
Purpose of the research: Development of an automated morphological analyzer.
Research methods: in solving the tasks in the work used analytical methods. The elements of modeling are used to build mathematical models that describe linguistic patterns, as well as methods of object-oriented programming.
Hardware: laptop Intel Core i3, Embarcadero RAD Studio XE3
Using the results of the study: an automated system of morphological analysis was introduced into the educational process of the Osh technological University after named M.M. Adyshev and the electronic library of the Osh State Pedagogical Institute as a module for information retrieval. As well as the program received positive feedback from experts of the national commission on the State language under the President of the Kyrgyz Republic.
Scope: The research results and the developed system can be used in machine translation systems, expert systems, training systems, as a basic module.

Көчкөнбаева Бүажар Осмоналиевна

КЫРГЫЗ ТИЛИ ҮЧҮН МОРФОЛОГИЯЛЫК АНАЛИЗАТОРДУН МОДЕЛДЕРИН ЖАНА АЛГОРИТМДЕРИН ИШТЕП ЧЫГУУ

Техника илимдеринин кандидаты окумуштуулук
даражасын изденип алуу үчүн жазылган диссертацияга

АВТОРЕФЕРАТ

Басылмага кол коюлган: 17.05.2019 ж.
Форматы 60x84/16. Көлөмү 1,5 б.т.
Офсеттик кагаз. Нускамасы 60 даана.

Н. Исанов ат. Кыргыз мамлекеттик курулуш,
транспорт жана архитектура университети
«Авангард» окуу-басма борбору
720020, Бишкек ш., Малдыбаев көч., 34, б