Л.Н. Гумилев атындағы Еуразия ұлттық университеті

ӘОЖ 004.912 Қолжазба құқығында

**БАТЫРХАНОВ АРДАК ГАБИТОВИЧ**

**Ғылыми-техникалық мәтіндерден терминологиялық ақпараттарды алу бағдарламалық құралдары мен әдістерін жасау**

6D070300 – Ақпараттық жүйелер

Философия докторы (PhD)

дәрежесін алу үшін дайындалған диссертация

Ғылыми кеңесшісі

физика-математика ғылымдарының

кандидаты, доцент м.а.

Бакиев М.Н.

Шетелдік ғылыми кеңесшісі:

техника ғылымдарының кандидаты

Загорулько Ю.А.

(НовоСібір)

Қазақстан Республикасы

Астана, 2023

**МАЗМҰНЫ**

|  |  |
| --- | --- |
| **НОРМАТИВТІ СІЛТЕМЕЛЕР**..................................................................... | 3 |
| **АНЫҚТАМАЛАР**........................................................................................... | 4 |
| **БЕЛГІЛЕУЛЕР МЕН ҚЫСҚАРТУЛАР**.................................................... | 7 |
| **КІРІСПЕ**............................................................................................................ | 8 |
| **1 МӘТІНДЕРДЕН ТЕРМИНДЕРДІ АЛУ ӘДІСТЕРІ**.............................. | 19 |
| 1.1 Терминдерді алудың статистикалық және лингвистикалық критерийлері..................................................................................................... | 19 |
| 1.2 Терминдерді алудың дәлдігін арттыру жолдары | 21 |
| 1.3 Ғылыми-техникалық мәтіндерден терминологиялық ақпарат сөздіктерін қалыптастыру әдістемесі.............................................................. | 23 |
| 1.4. Шығарылатын құралымдардың ресми сипаттамасы............................... | 26 |
| **2 ТЕРМИНДЕРДІ ҚОЛДАНУДЫҢ ЛЕКСИКАЛЫҚ-СИНТАКСИСТІК ҚАЛЫПТАРЫ**............................................................... | 30 |
| 2.1 Ғылыми-техникалық терминдердің ерекшеліктері.................................. | 30 |
| 2.2 Терминдердің құралымы мен қолданылу нұсқалары.............................. | 32 |
| 2.3 Лексика-синтаксистік қалыптардың құрамы............................................ | 36 |
| 2.4 Құралым және терминдер нұсқаларының қалыптары | 38 |
| **3 ТЕРМИНДЕРДІ ТҰТЫНУДЫ АЛУДЫҢ РӘСІМ**................................. | 44 |
| 3.1 Рәсімдердің жұмыс схемасы....................................................................... | 44 |
| 3.2. Мәтінге қалыптарды қабаттастыру ережелері......................................... | 49 |
| 3.3 Сөздіктік және сөздіктік емес терминдерді алу........................................ | 53 |
| 3.4 Авторлық терминдер мен терминдерді қосылыстардан алу.................... | 56 |
| 3.5 Терминдердің синонимдерін алу................................................................ | 59 |
| 3.6 Терминдердің мәтіндік нұсқаларын тану.................................................. | 61 |
| 3.7 Алу рәсімдерінің бағдарламалық қамтамасыз ету.................................... | 63 |
| **4 АЛУ РӘСІМДЕРІН БІРІКТІРУ СТРАТЕГИЯСЫ**................................ | 66 |
| 4.1 Рәсімдердің жұмысын эксперименттік зерттеу......................................... | 66 |
| 4.2 Зерттеу нәтижелерін талдау........................................................................ | 67 |
| 4.3 Рәсімдердің нәтижелерін біріктіру стратегиясы....................................... | 69 |
| 4.4 Қолданбалы тапсырмаларда алу рәсімдерін қолданылуы........................ | 78 |
| **ҚОРЫТЫНДЫ**................................................................................................ | 82 |
| **ПАЙДАЛАНЫЛҒАН ӘДЕБИЕТТЕР ТІЗІМІ**.......................................... | 83 |
| **ҚОСЫМША А** – Лексикалық-синтаксистік қалыптар тілінің синтаксисі........................................................................................................... | 92 |
| **ҚОСЫМША Ә –** Терминдерді анықтау қалыптары және олардың синонимдерін ендіру......................................................................................... | 95 |
| **ҚОСЫМША Б –** Сөздіктік қалыптар мысалдары....................................... | 98 |
| **ҚОСЫМША В –** Терминдердің қолданылу нұсқаларының қалыптары... | 100 |
| **ҚОСЫМША Г –** Стратегия ережелері.......................................................... | 102 |
| **ҚОСЫМША Ғ –** Авторлық куәлік................................................................. | 104 |
| **ҚОСЫМША Д –** Ендіру актісі........................................................................ | 105 |

# **НОРМАТИВТІ СІЛТЕМЕЛЕР**

Диссертациялық жұмыста келесі стандарттарға сілтемелер қолданылған:

ҚР СТ 34.005-2002. Ақпараттық технология. Негізгі терминдер мен анықтамалар.

ҚР СТ 34.019-2002. Ақпараттық технология. Бағдарламалық құралдардың өмірлік циклінің процестері.

ҚР СТ 34.019-2005 (ISO/IEC 12207:1995, MOD). Ақпараттық технология. Бағдарламалық құралдардың өмірлік циклінің үрдістері.

ҚР СТ 34.002-2004. Ақпараттық технология. Есептеу электронды дара машиналар. Құрамына және сапа сипаттамаларын бағалау ережелеріне талаптар.

ГОСТ Р 53622-2009. Ақпараттық технологиялар. Ақпараттық-есептеу жүйелері. Өмірлік циклдің кезеңдері мен кезеңдері, құжаттардың түрлері мен толықтығы. – Тарм. 01.01.2002.

ГОСТ Р 7.0.66-2010. Бойынша стандарттар жүйесі ақпарат, кітапхана және баспа ісі. Құжаттарды индекстеу. Координаттық индекстеуге қойылатын жалпы талаптар.

ISO 25964-1:2011. Ақпарат және құжаттама – Тезаурустар және өзге сөздіктермен үйлесімділігі – 1 бөлім: Ақпараттық іздеуге арналған тезаурустер.

ISO 25964-1:2013. Ақпарат және құжаттама – Тезаурустар және өзге сөздіктермен үйлесімділігі – 2 бөлім: Өзге сөздіктермен өзара байланысы.

# **АНЫҚТАМАЛАР**

Бұл диссертaциялық жұмыста келесі терминдерге сәйкес aнықтaмaлaр қолдaнылған.

**Мәтіннен тыс критерийлер (off-page, внестраничные парақтан тыс)** - құжаттың мәтінінде қарастырылмаған және құжаттан ешқандай жолмен туындамайтын факторларды қамтитын іздеу жүйелерінің құжаттарды ранжирлеу критерийлері.

**Ранжирлеу** - өзінің сұрағына іздеу жүйесінен жауап таба отырып нәтижесін көру процесі.

**Инвертирленген файл (inverted file, инверсный файл, инвертированный индекс, инвертированный список, инверстік файл, инвертирленген индекс, инвертирленген тізім)** - құжаттар жинағының сөздерін тізімдейтін және әр сөз үшін оның барлық орындары көрсетілген іздеу жүйесінің индексі.

**Индекс (index, указатель/көрсеткіш)**- қараңыз Индекстеу.

**Дәйексөздер индексі (citationindex, цитирований)** –дәстүрлі библиографиялық ғылымдағы ғылыми мақаланың сілтемелер саны (дәйексөздер) мысалы бір жыл уақыт аралығында есептеледі.

**Индекстеу (indexing, индексация)** - құрастыру процесі немесе көрсеткішті (индекс) анықтау/приписывания процесі - кейін іздеуге қажет мәліметтер құрылымы.

**Ақпаратты іздеу (Information Retrieval, IR)**- құрылымдалмаған ақпаратты іздеу, оны ұсыну бірлігі еркін форматтағы құжат болып табылады. Сұраныс бойынша бейресми көрсетілген пайдаланушының ақпараттық қажеттілігі. Іздеу критерийлері және нәтижелері анықталмаған. Осы белгілермен ақпараттық іздеу «деректерді іздеуден» ерекшеленеді: ол формальды түрде анықталған предикаттар жиынтығымен жұмыс істейді және нәтижесі әрдайым анықталатын құрылымдалған ақпаратпен айналысады. Ақпаратты іздеу теориясы іздеу процесінің барлық құрамдас бөліктерін, атап айтқанда мәтінді алдын-ала өңдеу (индекстеу), сұраныстарды өңдеу және орындау, рейтинг, қолданушы интерфейсі мен кері байланысты зерттейді.

**Клоакинг (cloaking)** - іздеу жүйелерінің роботын тануынан, ол үшін арнайы мазмұнды генерациясынан құралған спамның іздеу техникасы ол пайдаланушыға көрсетілген мазмұннан түбегейлі ерекшеленеді.

**Латентно-семантическое индексирование** - факторлық талдаумен бірдей мағынасында патенттелген іздеу алгоритмі. Ол сөздердің құжаттармен байланысы матрицасының сингулярлы ыдырауына негізделген.

**Лемматизация (lemmatization, нормализация, қалыпқа келтіру)** - сөз тұлғасын сөздік формасына, яғни леммаға дейін қысқарту.

**Құжаттардағы кері жағдайлар (inverted document frequency, IDF, обратная частота в документах, обратная документная частота, құжаттардағы кері жиілік,)** - сөздің іздеу мәнінің көрсеткіші (оның айрықша күші); **«**обратная/кері» дейді, себебі осы көрсеткішті есептеу кезінде бөлшектің бөлімінде көбінесе берілген сөзден тұратын құжаттар саны тұрады.

**Кері байланыс** - іздеу нәтижесіне пайдаланушылардың жауабы және олардың табылған құжаттардың релеванттылығы туралы пайымдаулары, іздеу жүйесіне сұранысты тез өзгерту үшін тіркеледі. Бірнеше алғашқы табылған құжаттар автоматты түрде релевантты болып есептелетін жалған кері байланыс - сұранысты модификациялау/өзгерту техникасынан ажырату қажет. Көптеген алғашқы табылған құжаттарды автоматты түрде релевантты деп тауып, сұранысты өзгерту техникаларынан жалған кері байланысты ажырата алу керек.

**Негіз** – сөздің бір бөлігі, сөзжасамдық және сөз өзгерту жиынтықтары үшін жалпы форма.

**Мағынасы бойынша іздеу** - сұраныс сөздері жоқ құжаттарды табуға қабілетті ақпаратты іздеу алгоритмі.

**Ұқсас құжаттарды іздеу (similar document search)** - ақпараттық іздестірудің міндеті, сұраныс ретінде құжаттың өзін қолдану арқылы оған максималды ұқсас құжаттарды табу.

**Іздестіру жүйесі** (search engine, SE, информационно-поисковая система, ИПС, поисковая машина, машина поиска, «поисковик», «искалка», ақпараттық іздестіру жүйесі, АІЖ, іздеу машинасы, “іздеуші”) - мәтіндік құжаттарды іздеуге арналған бағдарлама.

**Іздеу тәртібі/предписание (query, запрос/сұраныс)** - әдеттегі мәтін жолы.

**Полисемия (polysemy, многозначность/көпмағыналылық)** - бір сөзде бірнеше мағынаның болуы.

**Толықтығы (recall, охват/қамту)** -жинақтағы барлық релевантты материалдарға қатысы бойынша іздеу жүйесіндегі жауапқа тиісті материалдардың үлесі.

**Көшірмеге жақын (near-duplicates, приблизительные дубликаты/шамамен көшірмелер)** - Көшірмелер.

**Прюнинг (pruning)** - іздеу кезінде сұранысты тездету мақсатында релевантты емес құжаттарды әдейі кесіп тастау.

**Тікелей іздеу** - алдын-ала өңдемей құжаттардың мәтіні бойынша тікелей іздеу.

**Сөздің ерекше күші (term specificity, term discriminating power, контрастность, различительная сила/ерекше күш)** - сөздің тарлық немесе кеңдік дәрежесі. Іздеудегі тым кең терминдер тым көп ақпарат береді, ал оның айтарлықтай бөлігі пайдасыз. Тым тар терминдер өте аз құжаттарды табуға көмектеседі, дегенмен олар нақтырақ болады.

**Релеванттылық (relevance, relevancy)** - сұраныстың құжатқа сәйкестілігі.

**Пертиненттік (пертинентный/Пертиненттілік, enPertinence, frPertinence)**- алынған ақпараттың ақпараттық қажеттілікке сәйкестігі.

**Сигнатура (signature, подпись/қол)** - Сигнатура әдісі бойынша зерттеу кезінде жинақтағы барлық блоктардың сигнатураларының сұраныс сөздерінің хэш мәндерімен сәйкестігін іздеуде дәйекті түрде қарастырылады.

**Сөз өзгерісі (inflection)** - осы типтегі сөздерге тән формалардың (парадигманың) тіркелген жиынтығына жататын белгілі бір грамматикалық мәтінде талап ететін нақты грамматикалық мағынасына қарай форманы қалыптастыру. Сөзжасамдықтан айырмашылығы, ол типтің өзгеруіне әкелмейді және болжанатын мәнді тудырмайды. Атаулардың сөз өзгерімін септеу (declension), ал етістіктің сөз өзгерімін жіктеу деп атайды.

**Сөзжасамдық/Словообразование (derivation)** – сөздің жасалуы немесе басқа сөздің негізі.

**Іздеу жүйесінің спамы (spam, спамдексинг, накрутка поисковых систем/іздеу жүйесін орамалдау)** - құжат авторларының тарапынан ақпаратты іздеу нәтежесіне әсер ету әрекеті.

**Стемминг** – негізгі сөзді бөліп алу үдерісі.

**Стоп-сөздер (stop-words)**- бұл іздеу жүйесі өзінің өнімділігін және іздеу нақтылығын арттыру үшін индекстеу және іздеу процесінен шығарылып тасталған жалғаулар, сөзалды қосымшалар және басқа да жиілік сөздер.

**Жұрнақ ағаштар, жұрнақ массивтері (suffix trees, suffix arrays, PAT-arrays, суффиксные деревья)** - «бор» (trie) деген терминмен белгілі деректер құрылымындағы барлық маңызды мәтін суффикстерінің ұсынылуына негізделген индекс. Мәтіннің белгілі бір позициясынан басталып оның соңына дейін жалғасатын кез-келген «ішкі жол» бұл индекстегі жұрнақтар. Нақты қосымшаларда жұрнақтардың ұзындығы шектеулі, тек маңызды позициялар индекстеледі - мысалы, сөздің басы. Бұл индекс инверттелген файлдарда жасалған индекске қарағанда күрделірек сұраныстарды орындауға мүмкіндік береді.

**Токенизация (tokenization, lexical analysis, графематикалық талдау, лексикалық талдау)** - мәтіндегі сөздерді, сандарды және басқа токе/белгілерді бөліп көрсету, мысалы, сөйлемдердің шекарасын табу.

**Нақтылық (precision)** - іздеу жүйесінің жауабындағы релевантты материалдың үлесі.

**Хеш-мағынасы (hash-value)** - хеш-функцияларының мағынасы (hash-function), еркін ұзындықтағы мәліметтерді (әдетте жолды) белгіленген тәртіп нөміріне түрлендіреді.

**Құжаттардағы жиілік (сөздер) (document frequency, құжаттарда кездесу, құжат жиілігі)** - берілген сөзді қамтитын жинақтағы құжаттар саны.

**Термин жиілігі (term frequency, TF)** - құжаттағы сөзді қолдану жиілігі.

**Шингл (shingle)** - тіркелген/фиксированной ұзындықтағы мәтін сөздерінің үздіксіз тізбегі хэш мәні.

**PageRank (Бет рейтингі)** - Интернеттегі парақтың статикалық (ғаламдық) танымалдылығын есептеу алгоритмі Лоуренс Пейждің есімімен аталады. Пайдаланушылардың кездейсоқ адасқан модельдер парағына кіру ықтималдығына сәйкес келеді.

**TF\*IDF** - сандық өлшеу векторлық модельдегідей сөз бен құжаттың сәйкестігі; көп болған сайын, соғұрлым құжатта сөздер жиірек кездеседі және жинақта салыстырмалы түрде азырақ болады.

**БЕЛГІЛЕУЛЕР МЕН ҚЫСҚАРТУЛАР**

|  |  |
| --- | --- |
| АТ | – Ақпараттық технологиялар |
| АЖ | – Ақпараттық жүйелер |
| МАӨ | – Автоматты түрді өңдеу |
| ТТ | – Табиғи тіл |
| ҒТ | – Ғылыми-техникалық |
| БНТ | – Бэкус-Наур тұлғалары |
| БҚ | – Бағдарламалық құралдар |
| ИмЕТ | – Информатика мен есептеу техникасы |
| ЭАР | – Электронды ақпараттық ресурстар |

**КІРІСПЕ**

**Зерттеу мәселесінің өзектілігі.** Ғылым мен техниканың қарқынды дамуы, есептеу кешендері мен компьютерлік желілердің кең таралуы, әртүрлі мәліметтер көлемінің көшкіндей ұлғаюы – осының бәрі есептеу жүйелері негізінде ақпаратты жинау, сақтау және өңдеу әдістерін әзірлеу мен жетілдіруді бұрынғыдан да өзекті ете түседі

Есептеу жүйелерімен өңделетін ақпараттың едәуір бөлігі адамдар үшін дәстүрлі түрде - табиғи тілдегі мәтіндер түрінде ұсынылған, бұл мәтінді автоматты түрді өңдеудің (МАӨ) бірқатар қолданбалы міндеттерінің бар болуын түсіндіреді. Оларға мыналар жатады:

* мәтіндерді бір табиғи тілден (ТТ) екіншісіне аудару;
* ақпаратты іздеу үшін құжаттар мен пайдаланушылардың сұраныстарын индекстеу;
* мәтіндердің жіктелуі мен айдарланып бөлінуі;
* мәтіндерді қысқаша мазмұндау және аңдатпалау;
* мәтін топтамаларынан ақпарат алу;
* пәнді саланың мәтіндері бойынша онтологияларды құру;
* компьютерлік сөздіктерді құру жән жаңарту [1].

Көрсетілген МАӨ қолданбалы есептерін шешу, жалпы алғанда графематикалық, морфологиялық, синтаксистік және семантикалық талдауды қамтитын табиғи тілдегі мәтінді (ТТ-мәтін) кезең-кезеңімен өңдеуді талап етеді [2]. Графематикалық талдау жасау кезінде мәтіндегі сөздер ерекшеленеді, мәтін сөйлемдерге, абзацтарға жіктеледі. Морфологиялық талдаудың көмегімен мәтіннің әр сөзі сөйлеудің қай бөлігіне жататындығы және оның қандай грамматикалық формада қолданылатындығы анықталады. Өз кезегінде, синтаксиситік талдау сөйлемдердің синтаксистік құрылымын анықтайды, ал семантикалық талдау жеке фрагменттер мен толықтай мәтіннің мағынасын анықтайды.

Көптеген қолданбалы есептерді шешкен кезде, жоғарыда аталған қадамдардың кейбірі жеңілдетілуі немесе тіпті төмендеуі мүмкін. Мысалы, құжаттарды индекстеу және рефераттау мәселелерін шешуде жиі өңделетін мәтіндерге толық синтаксистік және семантикалық талдау жүргізілмейді [3].

Талдау кезеңін жеңілдетуге тырысудың негізгі себебі оны жүргізудің күрделілігіне байланысты. Осы кезеңді орындау барысында, әдетте, тіпті салыстырмалы түрде аз сөйлем үшін ТТ-мәтін оны синтаксистік талдаудың бірнеше нұсқаларын жасайды. Сөйлем ұзындығының ұлғаюымен нұсқалардың саны экспоненциалды түрде артады, осылайша үлкен мәтіндер үшін олардың толық талдау уақыты мен пайдаланылатын машина ресурстарының көлемі көбінесе қолайсыз болып шығады. Сондықтан, МАӨ бірқатар міндеттерін шешу кезінде мәтіннің әр сөйлемін толық талдаудың орнына ішінара талдау жасалады, онда тек белгілі бір синтаксистік құралымдар, ең алдымен, сөз тіркестері танылады [4].

Ішінара синтаксистік талдауды қолдану онда өңделетін мәтіннің мазмұнын және олардың арасындағы байланысты көрсететін кілт сөздер мен сөз тіркестерін анықтауды қажет ететін МАӨ міндеттерін шешуде орынды болып шықты. Мұндай міндеттердің қатарына мәтіндерді индекстеу, жіктеу және айдарлау жатады [5]. Сондай-ақ, ішінара синтаксистік талдау ТТ-мәтіндерден ақпарат алу кезінде қолданылады (нақтырақ, аталған нысандарды: есімнаманы, мекенжайларды, географиялық атауларды, тауар атауларын алу кезінде), оған қоса, белгілі бір пәндік саланың (БҚ) мәтіндері бойынша машиналық сөздіктерді құру және жаңарту кезінде де қолданылады [6]. Сонымен қатар, қажетті мәтіндік бірліктерді неғұрлым толық және дәл тану үшін көбінесе олардың құрылымы ғана емес, сонымен бірге оларды өңделген ТТ-мәтіндерде қолдану ерекшеліктері де ескеріледі.

Лингвистикалық зерттеулер көрсеткендей, ТТ-мәтіндерінің және оларда қолданылатын бірліктердің маңызды ерекшеліктері, ең алдымен, сөйлеудің әртүрлі функционалды стильдерімен байланысты [7], олардың пайда болуы адамдардың қарым-қатынасының мақсаттары мен шарттарының әртүрлілігіне байланысты. Әдетте ауызекі, публицистикалық, ғылыми-техникалық стильдер мен көркем әдебиет стилі ерекшеленеді. Стильдер арасындағы айырмашылық тілдің барлық деңгейлерінде байқалады, бірақ ол әсіресе лексикалық және синтаксистік деңгейлерде айқын көрінеді: мәтіндердің рұқсат етілген сөздік құрамы және оларда қолданылатын синтаксистік құралымдар белгілі бір стильге байланысты [8].

Ғылыми-техникалық стиль ең айқын ерекшелікке ие. Оның сипаттамалық ерекшеліктерінің бірі - мәтіндердің терминдермен, яғни белгілі бір БҚ ұғымдарын атайтын сөздер мен сөз тіркестерімен қанықтылығы (*коммуналдық құрылыстар, спектрлік сәулелену коэффициенті, сыртқы құрылғыдан үзіліс және т.б.*) [9]. Терминдер әдетте ғылыми-техникалық мәтіннің (ҒТ- мәтіні) ең жиі бірліктерінің қатарына кіретіндіктен және оның мазмұнын дәл көрсететіндіктен, оларды тану МАӨ қолданбалы міндеттерінің көпшілігін шешуде талап етіледі. Қазіргі МАӨ жүйелерінде терминологиялық сөздер мен сөз тіркестерін алу көбінесе ішінара синтаксистік талдау негізінде жүзеге асырылатындығын ескереміз.

БҚ терминдері және онымен аталатын ұғымдар ғылыми зерттеулер барысында пайда болады және тиісті ҒТ-мәтіндерінде (баяндамалар, мақалалар, монографиялар), содан кейін осы мәтіндер негізінде құрастырылған сөздіктер мен оқулықтарда тіркеліп жазылады. ҒТ мәтіндерде кездесетін терминологиялық сөздер мен сөз тіркестерінің арасында әдетте [10]:

* жалпыға ортақ терминдер, олар әдетте қолданыстағы терминологиялық сөздіктерде жазылады;
* сипатталған ғылыми зерттеулер барысында пайда болған ұғымдарды белгілеу үшін енгізілген жаңа терминдер - олар әдетте сөздіктерде жоқ.

ТТ-мәтіндерден жалпыға ортақ терминдерді анықтау үшін терминдердің компьютерлік сөздіктеріне негізделген автоматтандырылған рәсімдер әзірленді және кеңі ауқымда қолданылады [11]. Егер мәтінде оның болуы мүмкін грамматикалық тұлғаларының бірінде қолданылатын сөз немесе сөз тіркесі сөздікке енсе, онда ол белгілі термин ретінде танылады.

Мәтіндерден жаңа терминдерді автоматты түрде алу үшін лингвистикалық және статистикалық критерийлерге негізделген эвристикалық әдістер қолданылады. Статистикалық критерийлер негізінен өңделетін мәтіндегі немесе мәтіндер топтамасындағы сөздердің жиілігін, сондай-ақ осы жиіліктер негізінде есептелетін статистикалық шамаларды пайдаланады . Оның ішінде, кейбір МАӨ жүйелерінде терминдер ретінде пайда болу жиілігі берілген шекті мәннен жоғары сөздер мен сөз тіркестері қарастырылады. Статистикалық критерийлер талданатын мәтіндердің көлемі неғұрлым көп болса, соғұрлым жақсы жұмыс істейтінін ескертеміз, сондықтан қазіргі МАӨ жүйелерінде бұл критерийлер мәтін топтамаларынан терминдерді алу үшін қолданылады [12].

Лингвистикалық критерийлер, ең алдымен, терминологиялық тіркестердің құрылымына сүйенеді. Терминдер негізінен бір, екі және үш сөзден тұратын номиналды тіркестер болғандықтан, мәтінде ұқсас құрылымы бар болып танылған тіркестер потенциалды терминдер ретінде қарастырылуы мүмкін. Терминдердің белгілі бір құрылымы әдетте терминді құрайтын сөздердің сөз таптарын және олардың арасындағы синтаксистік байланыстарды анықтайтын **синтаксистік үлгі** ретінде сипатталады. Мысалы, N- бір сөзден тұратын терминдерді сипаттайтын үлгі - зат есімдер (*вектор, аорта*), ал N- сын есімнен және одан кейінгі зат есімнен тұратын екі сөзден тұратын терминдердің үлгісі (*тұжырымдамалық операция, маңызды мысал*), NNgen - зат есім және ілік септіктегі зат есімнен құралған терминологиялық тіркестерінің үлгісі (*жалпылық кванторы, рекурсия анафорасы*) және т. б.

Терминдер анықталған кезде, олардың құрылымынан басқа, терминдер қолданылатын тілдік құралымдар жиі ескеріледі. Қарастырылып отырған БҚ-ға тән құралымдарға сүйене отырып, мәтіннен мүмкін терминдер болып саналатын сөздер мен сөз тіркестері шығарылады [13]. Мысалы, жұмыста [14] ауылшаруашылық тақырыбындағы мәтіндерде түрдің құрылымдары танылады yieldsofSPECIES(yieldofwheat, yieldofrice - *бидай дақылдары, күріш дақылдары*), олардан SPECIES - өсірілген дақылдардың атаулары шығарылады.

Статистикалық және лингвистикалық критерийлерге негізделген мәтіндерден терминологиялық ақпаратты автоматты түрде алудың негізгі мәселелерінің бірі осы критерийлерге тек БҚ терминдері ғана емес, сонымен қатар жалпылама лексиканың кейбір сөздері мен сөз тіркестері де сәйкес келетіндігімен байланысты. Мысалы, ҒТ мәтіндерінде жиі кездесетін сөз тіркестері (олардың құрылымы жоғарыда келтірілген синтаксистік үлгілерге сәйкес келеді): *тапсырма, негізгі идея*, *ережені қолдану* (термин болмаса да) терминдер ретінде танылуы мүмкін. Сонымен қатар, бір сөйлемше бір пәндік салада термин болса, бірақ басқа сала мәтіндерінде термин бола алмайды: мысалы, *осы шақ* сөйлемшесі лингвистикалық мәтіндердегі термин, бірақ биология немесе информатика саласындағы мәтіндерде термин емес. Осылайша, ТТ мәтіндерінен терминдерді автоматты түрде алудың заманауи әдістері олардың жұмысының нәтижесінде тек ықтимал терминдерді немесе **үміткер терминдерді** алуға мүмкіндік береді, яғни, мұндай сөздер мен сөз тіркестері, олар үшін тек белгілі бір дәрежеде (дәлдікпен) терминдер деп айтуға болады.

Терминдерді алудың дәлдігі мәселесімен қатар, оларды танудың толықтығы мәселесі де сөзсіз туындайды. Статистикалық немесе лингвистикалық критерийлер терминдердің барлық ерекшеліктерін толық ескере алмайтындықтан, мәтіндерде стандартты емес синтаксистік құрылымы бар терминдер немесе төмен жиілікті терминдер жиі танылмайды.

Қазіргі МАӨ жүйелерінде терминдерді алудың толықтығы мен дәлдігін арттырудың негізгі әдісі статистикалық және лингвистикалық критерийдің қажетті комбинациясын таңдау болып табылады [15], бірақ бұл жағдайда да өңделетін мәтіндердің көлеміне және қолданылатын критерийлер комбинациясына байланысты терминдерді анықтау дәлдігі көбінесе 20%-дан 50% -ға дейін, ал анықтау толықтығы 55%-дан 85%-ға дейін өзгереді [16]. Осылайша, терминдерді алудың автоматты әдістерінің толықтығын және де, әсіресе, дәлдігін арттыру мәселесі әлі де өзекті болып қала береді.

Компьютерлік сөздіктерді құру және жаңарту немесе БҚ мәтіндері бойынша онтологияларды құру сияқты МАӨ қолданбалы міндеттерін шешу кезінде мәтіндердің үлкен топтамаларын өңдеу кезінде терминдерді алудың толықтығы мен дәлдігінің қолайлы мәндеріне қол жеткізіледі [17]. Сонымен қатар, МАӨ-дің көптеген өзге міндеттерінде статистикалық критерийлердің мүмкіндіктері айтарлықтай шектелген жеке алынған ҒТ мәтінінің терминдерін талдау қажет. Мұндай міндеттерге мыналар жатады: мәтінді бір ТТ-ден екіншісіне автоматты түрде аудару, мәтінді қысқаша мамұндау және аңдатпалау, құжаттың глоссарийі мен пәндік көрсеткішін жасау, сондай-ақ, ҒТ-мәтінді әдеби-ғылыми редакциялауды автоматтандыру [18]. Мұндай міндеттерде әртүрлі терминдерді ғана емес, сонымен бірге олардың терминдерді қолдану жиілігін ілеспе санаумен бірге талданатын мәтінге енуін де мүмкіндігінше толық тану қажет, бұл оның тұжырымдамалық мазмұнын неғұрлым дұрыс бағалауға мүмкіндік береді. Мәтіндегі терминдердің әртүрлі көріністерін анықтаудағы қиындықтар, ең алдымен, терминдерді қолданған кезде жиі өзгертілетіндігімен байланысты - қиылған, қысқартылған, синонимдермен ауыстырылған және т.б. [19]: *сұраныстың коммуникативті көп мәндігі - коммуникативті көп мәндік және синтаксистік ұғым - СинП, кірістірілген файл - тіркеме*. Мұндай мәтіндік нұсқалар бір ұғымның әр түрлі өрнек тұлғалары болып табылады және мүмкіндігінше мәтінді өңдеу кезінде танылуы керек. Мысалы, мәтінде [20] *ақпараттық жүйе* термині 32 рет қолданылады, оның 7-і бастапқы түрінде, 4 рет оның қиылған нұсқасы *жүйе* түрінде, ал 21 рет қысқартылған синонимі *АЖ* болып кездеседі. Терминдерді автоматты түрде алудың қазіргі заманғы әдістерінің көпшілігі мұндай мәтіндік нұсқаларды танымайды және олар есептеген *ақпараттық жүйе* терминін қолдану жиілігі 32 емес, 7 болады.

Жоғарыда көрсетілген мәтіндік нұсқалардан басқа, ҒТ мәтіндерінде бірнеше терминологиялық тіркестердің **қосылыстары** (комбинациялары) да кездеседі [21]. Терминдерді біріктірудің типтік мысалы - *кіріс және шығыс деректері,* екі терминнен құралған: *кіріс деректері және шығыс деректері*. Мұндай қосылыстар сонымен қатар МАӨ қолданбалы міндеттерін шешу кезінде ескеру қажет терминдердің мәтінге ену нұсқалары болып табылады.

Мәтіндік нұсқалар мен терминдердің қосылыстарын автоматты түрде тану үшін екі негізгі тәсіл қолданылады [22]. Бірінші тәсілдің бір бөлігі ретінде термин және оның болжамды нұсқасы (сөз немесе сөз тіркесі) таңбалар тізбегі ретінде қарастырылады және олардың ұқсастық дәрежесі арнайы эвристика арқылы сандық түрде бағаланады, содан кейін алынған сандық мәнге сүйене отырып, қарастырылып отырған сөздің (сөз тіркесінің) терминнің нұсқасы екендігі туралы қорытынды жасалады.

Екінші тәсіл аясында мәтіндегі терминдердің әртүрлі көріністерін анықтау үшін оларды түрлендіру ережелері қолданылады. Бұл ережелер терминнің әрбір синтаксистік үлгісі үшін жеке сипатталған [23]. Мысалы, түр ережесі А N -> А А N түрдің ағылшынша терминдерінің түрленуін сипаттайды А N (сын есім және одан кейінгі зат есім), және, атап айтқанда, acidic protein {*қышқыл ақуыз*) термині үшін мәтінде оның *acidic epidermal protein (эпидермистің қышқыл ақуызы)* нұсқасын тануға мүмкіндік береді. Жұмыста [24] осындай ережелерді қолдана отырып, мәтіндер кешенінде белгілі бір ұғымды білдірудің әртүрлі тәсілдерін іздеу мәселесі шешіледі.

Терминдердің нұсқаларын анықтаудың қарастырылған тәсілдерінің біріншісін тілдік тәуелсіз деп санауға болады, бірақ оның шеңберінде нұсқалардың барлық түрлерін тану мүмкін емес. Атап айтқанда, оның көмегімен *ноль - нөл, colour - color* түрінің мәтіндік нұсқалары сәтті танылады, алайда терминдік сөздер мен олардың арасындағы синтаксистік байланыстар өзгеретін нұсқаларды анықтау үшін (*оқу құрылғылары- оқитын құрылғылар*), бұл тәсіл оларды танудың толықтығы мен дәлдігінің қолайлы мәндерін бермейді.

Екінші тәсіл әртүрлі нұсқалардың көбірек түрлерін (соның ішінде көп сөзді терминдердің қосылыстарын) анықтауға мүмкіндік береді. Алайда, бұл тілге тәуелді, өйткені жалпы терминдердің түрлену ережелері оның нақты ТТ түріне байланысты. Осы уақытқа дейін бірқатар типтік мәтіндік нұсқалар мен терминдердің қосылыстарын құру ережелері ағылшын және француз тілдеріне арналған МАӨ жүйелерінде тұжырымдалған және қолданылған [25]. Олар ресми түрде жазылған және осы жүйеде қолданылатын терминдер ТТ мәтінінен және олардың нұсқаларынан шығарылған рәсімдерден бөлінген, бұл түрленудің жаңа жағдайларын тану үшін қолданыстағы ережелер жиынтығын өзгертуге және кеңейтуге мүмкіндік береді. Қазақ тілі үшін ұқсас ережелер ресімделмеген және терминологиялық нұсқаларды анықтаудың тиісті рәсімдері орындалмайды.

Осылайша, терминдерді автоматты түрде алу және олардың нақты қолданбалы тапсырма үшін қолайлы толықтығы мен дәлдігі бар талданатын мәтіндерге әртүрлі кіруі мәселесі әлі де өзекті болып қала береді.

Бұл диссертация қазақ тіліндегі берілген ҒТ мәтінінен терминологиялық ақпаратты автоматты түрде алу міндетін қарастырады:

* жалпыға ортақ терминдер;
* жаңа терминдердің анықтамаларын жобалау және олардың синонимдерін енгізу;
* танылған терминдердің мәтіндік нұсқалары;
* терминологиялық сөз тіркестерінің қосылуы;
* мәтінде танылған терминдер мен нұсқаларды қолдану жиілігі.

ҒТ мәтінінің көлемі аз болуы мүмкін (ғылыми мақалалар, аңдатпалар) және статистикалық критерийлер тек айтарлықтай көлемдегі мәтіндер үшін жақсы жұмыс істейтіндіктен, мәселені шешуде лингвистикалық критерийлерді қолдануға баса назар аударылады.

Диссертацияда терминдерді, құралым және терминдерді қолдану нұсқаларын біркелкі ресімдеуге және осы формалдау негізінде өңделген ҒТ мәтінінен көрсетілген терминологиялық ақпаратты алудың автоматты рәсімдерін құруға мүмкіндік беретін тәсіл ұсынылған. Осыған қоса, рәсімдер әр түрлі ақпаратты (терминдердің анықтамалары, олардың мәтіндік нұсқалары және т.б.) шығаратындықтан, олардың жұмыс нәтижелерін біріктіру болжалды түрде ҒТ мәтіндерінен терминдерді алудың дәлдігі мен толықтығын жақсартуы керек. Атап айтқанда, қазақ тіліндегі ҒТ мәтіндеріне тән терминологиялық нұсқаларды (мәтінге терминдердің тікелей енуімен бірге) есепке алу, жалпы жағдайда, мәтіндегі терминдердің әртүрлі қолданылуын анықтаудың толықтығын арттыруы керек.

Терминдерді, сондай-ақ құралымдар мен оларды қолдану нұсқаларын алуда қолданылатын лингвистикалық ақпарат ресми **лексикалық-синтаксистік қалыптар** түрінде жазылады. Лексикалық-синтаксистік қалып сипатталған тілдік құралым сөздерінің құрамы мен синтаксистік байланыстарын (грамматикалық сәйкестендіру) бекітуге мүмкіндік береді: терминологиялық сөйлемше, терминнің анықтамалары (*Т- дегенді біз D деп түсінеміз*), оның синонимін енгізу (*Т1-ді Т2 деп те атаймыз*), мәтіндік нұсқа, терминдердің қосылуы және т. б.

Ұсынылған тәсіл шеңберінде ғылыми прозаға тән лексикалық-синтаксистік қалыптар түрінде жазылады:

* терминологиялық сөздер мен сөз тіркестерінің синтаксистік үлгілері;
* терминдердің анықтамаларын жобалау және олардың синонимдерін енгізу, сондай-ақ олардан терминдерді алу ережелері;
* терминдердің лексикалық-синтаксистік нұсқаларын құру ережелері (олар терминдерге кіретін сөздер мен олардың арасындағы синтаксистік байланыстарды өзгертеді);
* терминдердің қосылыстарының құрылымы және олардан терминологиялық сөздер мен сөз тіркестерін алу ережелері.

Қазақ тіліндегі берілген мәтіннен терминологиялық ақпаратты автоматты түрде алу үшін лексикалық-синтаксистік заңдылықтар түрінде танылатын тілдік құралымдардың ресми сипаттамасына негізделген рәсімдер жасалды. Рәсімдердің тиімділігі екі пәндік салаға қатысты мәтіндерде зерттелген: информатика мен есептеу техникасы (ИмЕТ) және физика. Ол үшін лексикалық-синтаксистік қалыптар түрінде ИмЕТ және физика сөздіктеріне енетін терминдер қосымша сипатталды [26], бұл рәсімдерді эксперименттік зерттеуде жалпы ғылыми прозаға да, атап айтқанда аталған ғылыми салаларға да тән тілдік құралымдардың үлгілерін қолдануға мүмкіндік берді. Зерттеу нәтижелері бойынша терминдер мен олардың әртүрлі қолданылуын анықтаудың толықтығы орта есеппен 77% құрады.

Басқа ғылыми салаларда осындай қалпына келтіру тиімділігіне қол жеткізу үшін осы ғылыми салаларға тән сөздік терминдерді, құралымдарды және терминдерді қолдану нұсқаларын ескеру (және қалып түрінде рәсімдеу) қажет екені анық. Алудың әзірленген рәсімдерінде қалыптар жиынтығының кірісі ретінде қолдану осы рәсімдерді әр түрлі БҚ мәтіндеріне сәйкес қолдану терминін қолданыстағы қалыптарды түзету және жаңа қалыптарды қосу арқылы өңдеу үшін жеткілікті түрде реттеуге мүмкіндік береді.

Жүргізілген зерттеу нәтижелері бойынша диссертацияда берілген ҒТ-мәтіннен терминдерді алудың дәлдігі мен толықтығы көрсеткіштерін жақсартуға, оның ішінде мәтінде оларды қолдану жиілігі туралы дәлірек ақпарат алуға мүмкіндік беретін рәсімдердің нәтижелерін біріктіру стратегиясы ұсынылған.

Диссертацияда берілген ҒТ мәтінінен терминологиялық ақпаратты шығаруға ұсынылған тәсіл және оның шеңберінде әзірленген рәсімдер МАӨ қолданбалы міндеттерін шешуде пайдаланыла алады, онда мәтіндегі терминдердің тұжырымдамалық мазмұнын ашу үшін мүмкіндігінше дәлірек және толық тану қажет. Мұндай міндеттерге ҒТ мәтіндерін қысқаша мазмұндау және аңдатпалау, құжаттарды индекстеу және ақпараттық іздеуге сұраныстарды кеңейту, глоссарий мен құжаттың пәндік индексін құру жатады. Терминологиялық ақпаратты алудың әзірленген әдістері мен бағдарламалық құралдары ҒТ-мәтінді неғұрлым егжей-тегжейлі терминологиялық талдауды, оның ішінде шығарылған терминдердің байланыстарын анықтауды талап ететін қолданбалы міндеттерде де қолданылуы мүмкін.

Лексикалық-синтаксистік қалыптар түрінде ҒТ мәтіндеріндегі терминдердің қолданылуы туралы лингвистикалық ақпараттың бірыңғай формалды сипаттамасы, олардың негізінде берілген мәтіннен терминдердің әр түрлі қолданылуын алудың бағдарламалық рәсімдерін құру принципі, сондай-ақ осы рәсімдердің нәтижелерін біріктіретін стратегияны тұжырымдау диссертацияда алынған нәтижелердің ғылыми жаңалығын ұсынады.

**Зерттеудің мақсаты мен міндеттері**. Диссертациялық жұмыстың мақсаты ғылыми-техникалық мәтіндерден мамандандырылған терминологияны және онымен байланысты ақпаратты тиімді және дәл алуды қамтамасыз ететін жаңа әдістерді, алгоритмдер мен бағдарламалық құралдарды құру болып табылады.

Осы мақсатқа жету үшін шешілуі керек негізгі міндеттерге мыналар жатады:

1. Терминологиялық ақпаратты алудың қолданыстағы әдістері мен тәсілдерін зерттеу, олардың артықшылықтары мен кемшіліктерін анықтау.
2. Мәтінді алдын-ала өңдеу, үлгіні тану, терминдерді жіктеу және шығару, семантикалық ақпаратты талдау үшін жаңа әдістер мен алгоритмдерді әзірлеу.
3. Әзірленген әдістер мен алгоритмдерді оқыту және тестілеу үшін ғылыми-техникалық мәтіндер корпусын құру.
4. Терминологиялық ақпаратты алуды оңтайландыру үшін машиналық оқыту әдістері мен табиғи тілді өңдеу алгоритмдерін қолдану.
5. Пайдаланушылардың терминологияны шығару жүйесімен ыңғайлы өзара әрекеттесуін қамтамасыз ететін бағдарламалық құралдар мен интерфейстерді әзірлеу.
6. Дәлдік, толықтық және F-өлшем көрсеткіштеріне негізделген әзірленген әдістер мен бағдарламалық құралдардың сапасын бағалау, сондай-ақ бар шешімдермен салыстыру.
7. Мәтіндік деректердің үлкен көлемімен жұмыс істеу үшін әзірленген әдістер мен бағдарламалық құралдардың ауқымдылығы мен өнімділігін қамтамасыз ету.

**Жұмыстағы зерттеу объектісі** ғылыми-техникалық мәтіндер және олардағы терминология болып табылады.

Мұндай мәтіндер ғылыми зерттеулер, технологиялық әзірлемелер, инновациялық жобалар және ғылыми-техникалық қызметтің басқа аспектілері туралы ақпаратты қамтитын мамандандырылған құжаттар болып табылады.

**Жұмыстағы зерттеу пәні** ғылыми-техникалық мәтіндерден терминологиялық ақпаратты автоматты түрде алуға және талдауға арналған әдістер, алгоритмдер және бағдарламалық құралдар болып табылады.

Атап айтқанда, зерттеу пәні:

* Терминдерді және олардың басқа терминдер мен ұғымдармен байланысын алу үшін морфологиялық, синтаксистік және семантикалық талдауды қоса алғанда, табиғи тілдегі мәтіндерді талдау және өңдеу әдістері.
* Мәтіндердегі терминологиялық құрылымдар мен заңдылықтарды тану және жіктеу үшін қолданылатын машиналық оқыту алгоритмдері мен әдістері.
* Терминологиялық құрылымдарды, терминдерді қолдану нұсқаларын және олардың арасындағы синтаксистік байланыстарды ресімдеуге арналған лексикалық-синтаксистік заңдылықтар мен ережелер.
* Терминологиялық ақпаратты автоматты түрде алуға, басқа жүйелермен интеграциялауға және пайдаланушыға ыңғайлы түрде ұсынуға арналған бағдарламалық құралдар мен құралдарды әзірлеу.

**Ғылыми жаңалығы және қорғауға шығарылады:**

1. Терминологиялық сөз тіркестерінің құрылымын және терминдерді қолдану нұсқаларын біркелкі ресімдеуге және осы формалдау негізінде өңделген ғылыми-техникалық мәтінінен көрсетілген терминологиялық ақпаратты алудың автоматты рәсімдерін құруға мүмкіндік беретін тәсіл ұсынылған. Бұл олардың синтаксистік құрылымы мен байланыстарын сипаттайтын формальды лексика-синтаксистік шаблондар жүйесін жасауға мүмкіндік береді.

2. Қазақ тіліндегі берілген мәтіннен терминологиялық ақпаратты автоматты түрде алу үшін лексикалық-синтаксистік заңдылықтар түрінде танылатын тілдік құралымдардың ресми сипаттамасына негізделген рәсімдер принцитері жасалды. Ұсынылған тәсіл мәтіндегі терминдердің әртүрлі қолданыстарын, соның ішінде терминдердің тікелей кездесуімен қатар, олардың синонимдерін, анықтамаларын және басқа да байланысты конструкцияларды толық тануға мүмкіндік береді.

3. Ғылыми-техникалық мәтіннен терминдерді алудың дәлдігі мен толықтығы көрсеткіштерін жақсартуға, оның ішінде мәтінде оларды қолдану жиілігі туралы дәлірек ақпарат алуға мүмкіндік беретін рәсімдердің нәтижелерін біріктіру стратегиясы ұсынылған. Бұл стратегия әртүрлі процедуралардың нәтижелерін біріктіру және талдау арқылы мәтіннен терминдерді алудың дәлдігі мен толықтығын жақсартады.

**Автордың жеке қосқан үлесі.**Терминологиялық ақпаратты алудың тиімділігі мен дәлдігіне қол жеткізу үшін автор ұсынылған әдістер мен бағдарламалық құралдардың қолданыстағы тәсілдер мен технологиялармен салыстырғанда тиімділігін бағалау үшін эксперименттер жүргізді. Автор жұмысының нәтижесі ғылыми-техникалық мәтіндерден терминологиялық ақпаратты автоматты түрде алудың жаңа әдістері мен құралдарын әзірледі, бұл ғылыми-техникалық мәтіндерді талдау мен түсіну процесін едәуір жақсартуға мүмкіндік береді.

**Зерттеудің практикалық маңыздылығы және зерттеу нәтижелерін іске асыру.**

«Ғылыми-техникалық мәтіндерден терминологиялық ақпараттарды алу бағдарламалық құралдары мен әдістерін жасау» зерттеуінің маңыздылығы келесі аспектілерде жатыр:

**Теориялық маңыздылығы:** зерттеу табиғи тілді өңдеу, машиналық оқыту алгоритмдері және ақпаратты алу әдістері туралы білімді тереңдетуге ықпал етеді. Әзірленген әдістер мен алгоритмдер қолданыстағы тәсілдерді одан әрі жетілдіруге және терминология мен мәтінді талдау саласында жаңа шешімдер жасауға негіз бола алады.

**Практикалық маңыздылығы:** зерттеу нәтижелерін интеллектуалды іздеу, автоматты түрде сілтеме жасау, мәліметтер базасын құру, құжаттарды кластерлеу және жіктеу және ғылыми зерттеулерді қолдау сияқты ғылыми-техникалық ақпаратты талдауға қатысты әртүрлі салаларда қолдануға болады.

**Зерттеу қорытындылары мен нәтижелерінің анықтығы, сенімділігі және негізділігі** халықаралық ғылыми және ғылыми-практикалық конференцияларда апробациямен; зерттеу нәтижелерін Ѕсорuѕ және Web of Science (ClarivateAnalytics) халықаралық ғылымиметриялық базаларында индекстелетін ғылыми журналдарда жариялаумен, енгізудің актісін (Қосымша А) және екі авторлық куәлікті (Қосымша Ә) алумен қамтамасыз етілген.

**Жұмыс нәтижелерін апробациялау**. Диссертациялық жұмыстың негізгі нәтижелері және зерттеулердің нәтижелері келесі шетелдік, халықаралық, республикалық ғылыми-практикалық конференцияларда баяндалып талқыланды:

* Профессорлар Т. Н. Бияров, Вальдемар Вуйциктің 70 жылдығына және профессор Е. Н. Әмірғалиевтің 60 жылдығына арналған «Информатика және қолданбалы математика» IV Халықаралық ғылыми-практикалық конференциясы» (Алматы, 2019. – 25-29 қыркүйек);
* Физика-математика ғылымдарының докторы, профессор Рамазанов Мұрат Ыбырайұлының 70 жылдық мерейтойына орайластырылған «Математика, механика жəне информатиканың теориялық қолданбалы мəселелері» Халықаралық ғылыми конференцияның материалдары: Академик Е.А.Бөкетов атындағы Қарағанды мемлекеттік университеті (Қарағанды, 2019. 12 – 13 маусым);
* «Қазақстандық мұнай: өткен, қазіргі және болашақ»: Қазақстандық мұнайдың 120 жылдығына арналған халықаралық ғылыми-практикалық конференция материалдарының жинағы (Атырау, 2019).

Жұмыс Л.Н. Гумилев атындағы Еуразия ұлттық университетінің «Ақпараттық жүйелер» кафедрасының отырысында талқыланды (11 маусым 2020).

**Диссертацияның негізгі ережелері келесі ғылыми жұмыстарда жарияланған.** Диссертация материалдары бойынша 10 жұмыс жарияланды: оның ішінде 2 (екеуі) ҚР ҒЖЖБМ Ғылым және жоғары білім саласындағы сапаны қамтамасыз ету комитетінің ғылыми қызметтің нәтижелерін жариялау үшін ұсынылатын басылымдарда жарияланған, 3 (үшеуі) мақала Scopus және Web of Science қорының деректері бойынша нөлдік емес импакт-факторы журналда бар, 5 (бесеуі) – халықаралық конференция еңбектерінде.

**Диссертацияның құрылымы мен көлемі.** Диссертациялық жұмыс қазақ тілінде жазылған, кіріспеден, бөлімшелерге бөлінген өзара байланысты төрт бөлімнен, қорытындыдан және пайдаланылған әдебиеттер тізімінен тұрады.

Бірінші бөлімде терминологиялық сөздер мен сөз тіркестерін ТТ мәтіндерінен алудың қолданыстағы әдістеріне шолу берілген.

Екінші бөлімде лексикалық-синтаксистік қалып ұғымы енгізіліп, ғылыми-техникалық прозаға тән тілдік құралымдардың жасалуы сипатталған.

Үшінші бөлімде терминдерді, құралымдарды және терминдерді қолдану нұсқаларын лексикалық-синтаксистік қалыптар түрінде формалды сипаттамаға сәйкес ҒТ мәтінінен алудың әзірленген рәсімдері талқыланады, сонымен қатар алу рәсімдерінің бағдарламалық іске асырылуы сипатталады.

Төртінші бөлімде жүзеге асырылған рәсімдер жұмысының тиімділігі талданады және ҒТ-мәтіннен терминдерді алу стратегиясы сипатталған, оның мақсаты - алудың толықтығы мен дәлдігінің үйлесімді орташа мәні ретінде есептелетін аралас көрсеткіш - F-шараны ұлғайту.

Қорытындыда диссертациялық зерттеудің нәтижелері тұжырымдалған.

А қосымшасында лексикалық-синтаксистік қалыптар тілінің формалды сипаттамасы қалыпты Бэкус-Наур тұлғалары (БНТ) түрінде берілген. Басқа қосымшаларда:

* терминдердің анықтамалық қалыптары және олардың синонимдерін енгізу –Ә қосымшасынан қараңыз;
* сөздік қалыптарының мысалдары - Б қосымшасынан қараңыз;
* лексикалық-синтаксистік нұсқаларды құру ережелері және терминдерді қосылыстардан алу ережелері – В қосымшасынан қараңыз.
* Г қосымшасында терминдерді алудың ұсынылған стратегиясының ережелерінің ресми жазбасы (теориялық және бірнеше операцияларды қолдана отырып) келтірілген.
* Ғ қосымшасында авторлық куәлік.
* Д қосымшасында ендіру актісі.

Жұмыс 92 бетте көрсетілген, 10 сурет, 10 кесте бар. Пайдаланылған дереккөздердің тізімі 85 атаулардан тұрады.

**Алғыс**

Автор ғылыми кеңесші, М.Н.Бакиевке ғылыми-техникалық мәтіндер корпусын құру алгоритмдерін шешуге және модельдерін зерттеуге, терминологиялық ақпараттарды алудың оңтайландыру жолдарын және машиналық оқыту әдістерін зерттеуге қойған мінддетерді шешуге ықпал еткені, көптеген пайдалы кеңестері үшін ерекше алғыс білдіреді.

Сондай-ақ, шетелдік ғылыми кеңесші Ю.А. Загорулькоға (РҒА СБ ИЖИ, Новосібір), оның қарым-қатынасы осы зерттеу тұжырымдамаларын дамытуға айтарлықтай әсер еткенін айта келе, адал және жан-жақты көмегі, жан-жақты қолдауы, кәсібилігі және жұмысқа деген қызығушылығы үшін алғыс білдіреді.

Сондай-ақ, автор РҒА СБ А.П.Ершов атындағы Информатика жүйелері институтына (Новосібір), РҒА СБ Есептеу технологиялары институтына (Новосібір) және Новосібір мемлекеттік университетіне мекемелердің деректер қоры негізінде ғылыми зерттеулер жүргізуге көмек көрсеткені үшін алғысын білдіреді.

# **1 МӘТІНДЕРДЕН ТЕРМИНДЕРДІ АЛУ ӘДІСТЕРІ**

# 1.1 Терминдерді алудың статистикалық және лингвистикалық критерийлері

Терминдерді алудың алғашқы әдістері ХХ ғасырдың 60-70 жылдарында, ақпараттық іздеуді автоматтандыру процесінде табиғи-тілдік құжаттарды автоматты түрде индекстеу міндеті туындаған кезде, оларды шешу үшін кілт сөздер мен сөз тіркестерін олардың мағынасын дәл көрсететін бірліктер ретінде тануды талап еткен кезде жасалды [27].

Бастапқыда бұл мәселені шешу үшін өңделген құжаттағы жеке сөздердің пайда болу жиілігін ескеретін статистикалық критерийлер ғана қолданылды. Мәтіннің ең ақпараттық бірліктері онда бірнеше рет қолдануға бейім деген болжамға сүйене отырып, өңделетін мәтіннің барлық әр түрлі сөздерінің ішінен ең жиі кездесетіндері таңдалды.

Терминдерді танудың бұл тәсілінің негізгі кемшілігі өте айқын: сөздерді тек олардың мәтінде пайда болу жиілігі бойынша таңдау терминдерді алудың қолайлы дәлдігіне қол жеткізуге мүмкіндік бермейді, өйткені маңызды емес жалпыға ортақ лексика (мысалы*, себеп, типтік, сондықтан* жәнет.б*.*) сөздері де жиі кездесуі мүмкін, олар. Тағы бір кемшілігі - тек жеке сөздерді қолдану статистикасына назар аудару терминдерді алудың қолайлы толықтығын қамтамасыз ете алмайды, өйткені терминдер көбінесе сөз тіркестері болып табылады.

80-90 жылдары терминдерді алудың толықтығын арттыру үшін мәтінде тек жеке сөздердің ғана емес, сөз тіркестерінің де қолдану жиілігін ескеретін статистикалық әдістер әзірленді [28]. Бұл әдістер сөздер бір-бірінің қасында 3-5 сөз арақашықтықта қаншалықты жиі кездессе (бұл қашықтық сөздер арасындағы синтаксистік және/немесе семантикалық байланыстың болуын көрсетеді), соғұрлым олар сөз тіркесін құрайды деген болжамға сүйенеді. Сондықтан мәтінді өңдеу барысында әр сөз жұбы үшін олардың арасындағы қашықтыққа байланысты олардың бірге пайда болу жиілігі туралы ақпарат жиналады және жиналған ақпарат негізінде бұл сөздер тіркесті құра ма, жоқ па деген болжам жасалады. Мысалы, ықтимал екі сөз тіркестері ретінде олардың арасындағы қашықтықтың ең аз таралуымен өзара кездесудің ең жоғары жиілігі бар сөз жұптарын таңдауға болады [29]. Әрі қарай, бұл жұптар тұрақты тіркестер ретінде қарастырылады, олардың арасында ықтималды терминдер ретінде, мысалы, ең жиі кездесетіндері таңдалады. Егер алынған тұрақты сөз жұптарын бірлік ретінде қарастыратын болсақ, онда осы әдістердің көмегімен мәтінде ұзын сөз тіркестерін (үш сөзді, төрт сөзді тіркестер және т.б.) тануға болатындығын ескереміз [30].

Тіркестерді есепке алу терминдерді алудың толықтығын едәуір арттыруға мүмкіндік берді, бірақ іс жүінде дәлдікке әсер етпейді. Бұл мәтін бірлігінің ақпараттылығы мен оны қолдану жиілігінің байланысы туралы болжам әрдайым дұрыс бола бермейтіндігіне байланысты, себебі терминдер де, жалпы сөздер және сөз тіркестері де (*тапсырма, үлкен мағына* және т.б.) жиі кездесуі мүмкін. Нәтижесінде, терминдерді алуда тек статистикалық критерийлерге сүйену қолайлы дәлдікке кепілдік бере алмайтынына көз жеткіземіз. Оған қоса, статистикалық әдістер төмен жиілікті терминдерге, сондай-ақ шағын көлемді мәтіндерге нашар қолданылады – терминдерді алудың толықтығы төмендейді. Осы себептерге байланысты қазіргі уақытта тек статистикалық критерийлерге негізделген әдістер, әдетте, жеке мәтінге емес, мәтіндер топтамасына (кешеніне) қатысты қолданылады [31].

Сонымен қатар, терминдерді алудың статистикалық әдістерінің артықшылығы - олар терминдер туралы қосымша (лингвистикалық) ақпаратты пайдаланбайды және бұл, атап айтқанда, ерікті синтаксистік құрылымы бар терминдерді тануға мүмкіндік береді.

80-90 жылдары лингвистикалық ақпаратқа негізделген терминологиялық сөздер мен сөз тіркестерін алу әдістерінің белсенді дамуы басталды [32]. Лексикалық құралдарды - терминологиялық сөздіктер мен тезаурустарды тарту осы құралдарда жазылған терминдердің өңделетін мәтіндерінде танудың толықтығын едәуір арттырды. Сөздіктерде жоқ терминдерге келетін болсақ, оларды тану кезінде қазіргі лингвистикалық әдістер әдетте терминдердің жалпы құрылымдық ерекшеліктерін де, оларды өңделген ТТ мәтіндерде қолдану ерекшеліктерін де ескереді.

Терминдер негізінен белгілі бір құрылымы бар зат есімдер мен сын есімдерден тұратын тіркестерден құралады, олар әдетте синтаксистік үлгі ретінде жазылады. Үлгі терминді құрайтын сөздердің сөз табын, кейде олардың морфологиялық сипаттамаларын және морфемалық құрылымын бекітеді [33]. Мысалы, N - бір сөзден тұратын терминдерді сипаттайтын синтаксистік үлгі - зат есімдер (*атом, gene*), ал N PrN - зат есімнен және одан кейінгі зат есім мен қосымшадан тұратын терминдерді сипаттайтын үлгі *(*кристал *кемістігі, callbyпате*). Терминдерді автоматты түрде алу жүйелерінде егер өңделетін мәтіннің кейбір сөзі немесе сөз тіркесі қарастырылып отырған синтаксистік үлгіге сәйкес келсе, онда ол термин болып саналады[34]. Кейбір БҚ (медицина, биология, химия) терминдерінің құрамына көбінесе морфемалық құрылымның айқын белгілері бар сөздер кіреді [35]. Осылайша, медициналық терминдер әдетте грек немесе латын түбірлерінің және аффикстердің көмегімен жасалады: *стенокардия (стенос - тар, кардия - жүрек), кардиология (кардио (кардиодан) - жүрекке қатысты, логия (логос) — ілім)*. Морфемалық құрылымның ерекшеліктерін ескеру терминдерді алудың толықтығы мен дәлдігін арттыруға мүмкіндік беретіндіктен, кейбір заманауи әдістер терминдерді құрайтын сөздердің типтік морфемалық құрылымы туралы ақпаратты қолданады. Мысалы, жұмыста [36] медициналық мәтіндерді өңдеу кезінде, егер мәтінде морфемалық құрылымы *кардио* префиксін қамтитын жаңа сөз кездессе, онда мұндай сөзді термин ретінде қарастыру мүмкін болады.

Терминологиялық сөздер мен сөз тіркестерін тану үшін қолданылатын басқа лингвистикалық ақпарат олардың белгілі бір жанрдағы және тар БҚ тән белгілі бір тілдік құралымдарда (контекстерде) қолданылуын ескереді [37]. Жалпы, мұндай құралымдарды оларға кіретін сөздері және осы сөздер мен терминдердің синтаксистік байланыстары арқылы сипаттауға болады. Мысалы, ауылшаруашылық мәтіндеріне тән құралымда *effectof* INFLUENCE, (*effectofdrough t- құрғақшылықтың әсері, effectofcold — аяздың әсері*) *effect* және *of* сөздерінен кейін бірденегінге әсер ететін факторды (INFLUENCE) атайтын сөз тұрады. Сондықтан сипатталған контексте пайда болған сөз немесе сөз тіркесі термин ретінде алынады [38].

Кейбір жанрлардың мәтіндеріне тән және БҚ тілдік құралымдарда бірнеше терминдік қолданыс болуы мүмкін, бұл терминдерді автоматты түрде анықтауда да қолданылады [39]. Мысалы, құралымның *suchТ1 asТ2* түрі *(suchcrimesasmoneylaundering - ақшаны жылыстату сияқты қылмыстар),* мұндағы Т1 және Т2 - терминдер, оларды тек-түр қатынасымен байланыстырады, сондықтан мұндай құралымда кездесетін тіркестер Т1 және Т2 үлкен ықтималдықпен терминдер деп санауға болады [40].

Тек лингвистикалық критерийлерге негізделген терминдерді алу әдістерінің басты артықшылығы - олар үшін мәтіннің мөлшері де, ондағы терминдерді қолдану жиілігі де маңызды емес. Терминдерді алудың толықтығы ескерілетін лингвистикалық ақпараттың толықтығына байланысты болуы олардың кемшіліктеріне жатады. Дәлдікке келетін болсақ, қолданылатын лингвистикалық критерийлер терминдер үшін бірегей болмағандықтан, олардың негізінде анықталған сөздер мен сөз тіркестерінің арасында БҚ терминдері ғана емес, сонымен қатар жалпы лексикаға тән тіркестер де болуы мүмкін. Мысалы, терминдерге тән синтаксистік үлгі A N өңделген мәтінде терминдерді: *ықтималдық қағазы, уақыт қатары* және термин емес тіркестер де: *шағын бөлік, негізгі кемшілік, маңызды міндет* анықтауға мүмкіндік береді.

Осылайша, терминдерді автоматты түрде алудың заманауи әдістері (статистикалық және лингвистикалық критерийлер негізінде) өз жұмысының нәтижесінде тек үміткер-терминдер - сөздер мен сөз тіркестерін алады, оларды тек белгілі бір дәрежеде ғана терминдер деп айтуға болады.

Терминдерді алудың дәлдігі мен толықтығын арттырудың негізгі әдісі - статистикалық және лингвистикалық критерийлерді біріктіріп қолдану, бұл олардың артықшылықтарын біріктіруге мүмкіндік береді [41]. Мысалы, статистикалық әдістердегі терминдерді анықтаудың толықтығын арттыру үшін терминологиялық сөздіктерді қолдануға болады, олар өңделген мәтіндерінде жалпы қабылданған терминдерді кездесу жиілігіне немесе сиректігіне қарамастан анықтауды қамтамасыз етеді [42]. Басқа тәсілдер терминдерді алудың дәлдігін немесе толықтығының бірін ғана жақсартуға бағытталған.

# 1.2 Терминдерді алудың дәлдігін арттыру жолдары

Терминдерді алудың дәлдігін арттырудың бір жолы үміткер-терминдерді (қандай да бір әдіспен анықталған) олардың ықтималдығы ең аздарын жою арқылы сүзу. Ол үшін тоқтату-сөздерінің тізімі жиі қолданылады [43], оған әдетте қызметтік сөздер (жалғаулар, есімдіктер, бөлшектер және т.б.) және қарастырылып отырған БҚ [44] бойынша терминдер болып табылмайтын басқа сөздер кіреді. Тоқтату-сөздерінің тізімін қолданудың ең қарапайым және кең таралған нұсқасы көптеген термин-үміткерлерден, толығымен солардан тұратын тоқтату-сөздері мен сөз тіркестерінен тұрады [45]. Алайда, берілген БҚ терминдері болып табылмайтын барлық сөздерді ескеру мүмкін емес; сонымен қатар, осылайша жасалған тоқтату-сөздерінің тізімі басқа БҚ-ға жарамсыз.

Бұл мәселенің шешімдерінің бірі ақпараттық-іздеу жүйесінде [46] мәтіндер кешені бойынша БҚ тезаурусының қалыптасуын автоматтандыру кезінде ұсынылды. Тоқтату-сөздері алдымен сын есім, есімше немесе зат есім болып табылмайтын барлық сөздер болып саналды. Алайда, зерттеулер көрсеткендей, терминдерді құру кезінде барлық сөздер бірдей әрекет етпейді [47]: мысалы, N N үлгісіне сәйкес келетін және зат есімді қамтитын тіркестер - *қамсыздандыру* (*сапаны қамсыздандыру, қамсыздандыру шарттары*) терминдер емес, сонымен бірге А N түрінің тіркестерінің бір бөлігі дәл осндай зат есіммен (*бағдарламалық қамсыздандыру, ақпараттық қамсыздандыру*) терминдер болып табылады. Сондықтан жүйедегі тоқтату-сөздерінің тізімімен қатар терминологиялық үйлесімділік деп аталатын сөздікті қолдану ұсынылды [48], оның әр элементінде (сөз немесе сөз тіркесі) «-» немесе «0» белгісі бар. «-» белгісі, мысалы, сандармен (*жүз* және т. б.), анықтамалық сөздермен (*көрсетілген* және т. б.) және баға-сөздерімен (*жақсы* және т. б.) байланысты, ал «0» белгісі *үлкен, көлем, құру* және т. б. сияқты сөздермен байланысты. Қосымша «+» белгісі сөздікке қатысты барлық жаңа сөздерге жатады.

Көрсетілген белгілер мәтіндерден таңдалған сөз тіркестерін белгілейтін ережелерде қолданылады. Мысалы, ереже

A (-) + N(-) = G(-)

егер G тіркесін құрайтын А сын есімі де, N зат есімі де «-» белгісіне ие болса, онда бұл тіркес үлкен ықтималдықпен термин емес және оған «-» белгісі беріледі (осылайша талданатын тіркес тезаурусқа енгізу мүмкіндігін қарастыру үшін сарапшыға ұсынылмайды). Сонымен қатар ережеге сәйкес

A (+) + N ( -) = G (+)

G тіркесі «-» белгісін алады және сарапшының қарауына ұсынылады.

Көптеген үміткер-терминдерді қысқартудың тағы бір тәсілі - оларды ретке келтіру, содан кейін ең ықтимал кандидаттарды таңдау. Үміткер-терминдерді ретке келтіру көбінесе белгілі бір функцияның мәндері негізінде жүзеге асырылады, мысалы, олардың мәтіндегі кездесу жиілігі [49], ал реттелген тізімнен функция мәндері белгіленген шектен жоғары немесе төмен болатын сөздер мен сөз тіркестері таңдалады.

Көптеген үміткерлерді қысқартудың осыған ұқсас әдісі, атап айтқанда, [50] жұмысында қолданылады, ол басқа терминдердің бөлігі болып табылмайтын термин-сөз тіркестерін (ең ұзын) шығаруға арналған. Мысалы, мәтінде *санды бейнелеу* тіркесі жиілігі - 1, *үлкен разрядтылық саны* тіркесі жиілігі - 1 және *сан* сөзі жиілігі - 3 болып кедеседі, оның ішінде 2 рет *сан* сөзі жоғарыда аталған тіркестерінде кездеседі (*санның көрінісі және үлкен разрядтылық саны*). Егер сіз тек кездесу жиілігіне сүйенетін болсаңыз, онда ықтимал термин - бұл *сан* сөзі, бірақ тіркемелерді ескере отырып, үш үміткердің де қолдану жиілігі бірдей.

Осындай жағдайларды, сондай-ақ мәтінде ұзын сөз тіркестерінің қысқа сөздерге қарағанда жиілігі төмен қолдануға бейім екенін ескеру үшін *С-Уа1ие* [51] функциясы ұсынылды:

онда

*a* – сөз (сөз тіркесі),

– оның ұзындығы,

*freq(a)* – мәтіндегі *a* кездесу жиілігі,

*Ta* – *a* бар мәтіннің барлық сөз тіркестерінің жиынтығы,

*P(Ta)* – құрамында *а* бар сөз тіркестерінің саны.

Егер *а* басқа терминдерге салынбаса, онда *C-Value(a)* мәні оның ұзындығына және кешендегі пайда болу жиілігіне ғана байланысты екенін көру оңай; егер *а* басқа терминдерге салынған болса, онда *C-Value(a*) мәні олардың жиілігінің орташа мәніне азаяды.

Бұл жұмыста [52] көптеген үміткер-терминдерді қысқарту екі кезеңде жүзеге асырылатынын ескертеміз. Алдымен көптеген үміткерлер тоқтату-сөздерінің тізімімен қысқарады, содан кейін оның *C-Value* мәндеріне негізделген ретке келтіріледі және *C-Value* мәні берілген шектен жоғары болып табылатын терминдерді таңдайды.

# 1.3 Ғылыми-техникалық мәтіндерден терминологиялық ақпарат сөздіктерін қалыптастыру әдістемесі

Бір терминнің мәтініне жеке терминдерді де, әртүрлі көріністерді де алудың толықтығын арттырудың негізгі тәсілі - мәтінде терминдерді қолдану кезінде сөзсіз пайда болатын және *ақпараттық ресурс - ақпараттық ресурс –ресурс – ЭАР* болып табылатын нұсқаларды ескеру.

ҒТ мәтіндерінде терминдер жиі түрі өзгертіледі - қиылады, қысқартылады, синонимдермен ауыстырылады, бір-бірімен байланысады. Қазіргі уақытта мәтіндерде кездесетін терминологиялық нұсқалардың бірнеше жіктелуі белгілі. Олардың әрқайсысы терминографияның теориялық мәселелерін шешу үшін немесе нақты қолданбалы есептер шеңберінде құрылды.

Жұмыста [53] ұсынылған ағылшын тілі терминдерінің нұсқаларын жіктеуді анағұрлым толық деп санауға болады, оның негізінде медициналық мәтіндердегі терминологиялық нұсқаларды анықтау жүзеге асырылды. Авторлар келесі нұсқаларды ажыратады:

* орфографиялық:*amino acid - amino-acid, NF-KB - NF-kb, tumour–tumor;*
* морфологиялық: *cancer-cancers, cellular gene - cell gene;*
* құрылымдық: *clones of human - human clones, cell in blood - cell from blood, adrenal glands and gonads;*
* лексикалық: *carcinoma-cancer, haemorrhage - blood loss;*
* қысқартылған сөз: *DNA-deoxyribonucleic acid.*

Терминологиялық нұсқаларды автоматты түрде анықтауға бағытталған тағы бір толық жіктеу жұмыста келтірілген. Жіктеу жоғары тиімді баск тілі үшін жүргізілді, сондықтан оған жоғарыда келтірілген жіктеумен салыстырғанда термин сөздерінің келісілген септелуі/жіктелуі нәтижесінде пайда болатын агглютинативті (жалғамалы) нұсқалар қосымша енгізілді (*циклдік процесс - циклдік процестің - циклдік процестері*).

Көрсетілген жіктеулерде терминологиялық нұсқаларды түрлерге бөлу олардың қалыптасу әдісіне сәйкес жүзеге асырылады, бұл нұсқаларды автоматты түрде анықтау мәселесін шешу үшін маңызды: мысалы, орфографиялық нұсқалар - бұл термин сөздерінің әріптік құрамын өзгерту арқылы жасалған нұсқалар, морфологиялық нұсқалар - бұл термин сөздерінің морфемалық құрамын өзгерту арқылы жасалған нұсқалар және т. б.

Терминологиялық нұсқаларды қалыптастырудың негізгі әдістері көптеген еуропалық тілдер үшін ұқсас, сондықтан белгілі бір бейімделу жағдайында жоғарыда қарастырылған жіктеулер қазақ тіліне де қолданылады. Қазіргі уақытта қазақ тілді мәтіндерде кездесетін терминдердің белгілі жіктелімдері [54] теориялық есептерді шешу шеңберінде ұсынылғанын, сондықтан терминдерді автоматты түрде алу жүйелерінде қолдану үшін нақтылауды қажет ететіндігін ескереміз.

Терминологиялық нұсқалардың жіктелуінің белгілі бір ортақтығымен олардың жасалу ережелері әдетте белгілі бір тілге байланысты болады. Мысалы, қазақ және ағылшын тілдеріне тән құрылымдық нұсқаларды қалыптастыру кезінде *мекенжайлық құрсым*– мекенжай құрсымы*, address bus-bus of address* сөз тіркесінің синтаксистік байланыстарын бір уақытта өзгерте отырып, сөйлеудің басқа бөлігінің түбірімен ауыстырылады, яғни терминнің синтаксистік үлгісі өзгереді. Қазақ тілі үшін А N бұл ретте N Ngen-ге, ал ағылшын тілі үшін - А N болса N of N-ге ауыстырылады.Сондықтан жасалу ережелерін құру әр ТТ үшін бөлек жүргізілуі керек.

Терминологиялық нұсқаларды автоматты түрде анықтау үшін қазіргі уақытта екі негізгі тәсіл қолданылады. Бірінші тәсіл шеңберінде екі ерікті сөздерді/сөз тіркестерін (А және В) салыстыру әдістері әзірленді, олар [55] таңбалар тізбегі немесе сөз тізімдері ретінде қарастырылады; сондай-ақ, А салыстыру нәтижелері бойынша В-ның терминологиялық нұсқасы болып табылатындығын анықтау критерийлері де ұсынылған. Бұл тәсіл жазылуы ұқсас (атап айтқанда, орфографиялық) мәтіндік нұсқаларды тану үшін жақсы қолданылады.

А және В таңбалар тізбегі ретінде салыстыру үшін көбінесе бұлыңғыр салыстыру әдісі қолданылады, оның мәні келесідей: А және В терминдерін білдіретін таңбалар тізбегі N-грамдарға бөлінеді - N іргелес таңбалардан тұратын жолдар (мысалы, N=3-тегі аймақ сөзі келесі түрде ұсынылады: айм, йма, мақ. N-грамның әрқайсысындағы А әрбір N-грамдағы В-мен салыстырылады, және N-грамның бірдей саны есептеледі. Әрі қарай, есептелген мөлшерге сүйене отырып, мәндері 0-ден 1-ге дейінгі ауқымда болатын А және В ұқсастық коэффициенті арнайы эвристикалық функцияның көмегімен есептеледі. Егер ортақ N-грамм жоқ болса, коэффициент 0-ге тең, ал егер барлық салыстырылатын N-граммдар сәйкес келсе, коэффициент 1-ге тең болады. Егер ұқсастық коэффициентінің мәні 0 немесе 1-ге тең болмаса, бірақ алдын-ала белгіленген шектен жоғары болса, онда А және В терминологиялық нұсқалар болып саналады. Мысалы, жұмыста [56] сипатталған әдіс негізінде сөз тіркестері *extracellular signal-regulated kinase1* және *extracellular regulated-kinase-1* биохимия және молекулалық биология саласынан N=3 кезінде бір терминді қолдану нұсқалары ретінде танылды.

Жұмыста [57] А және В сөз тіркестерін салыстыру үшін оларға кіретін сөздерді жұптық салыстыру әдісі қолданылады. Сонымен қатар, салыстырылатын сөздердің әр жұбы үшін өңдеу қашықтығы деп аталады, яғни бір сөзді екіншісіне түрлендіру үшін енгізу/жою/ауыстыру қажет таңбалар саны есептеледі (осылайша *тэг* сөзін бір таңбаны ауыстыру арқылы *тег* сөзіне түрлендіруге болады: *э*-ден *е*-ге). Егер олар үшін өңдеу қашықтығының мәні берілген шекті мәннен төмен болса, екі сөз эквивалентті болып саналады *p* (мысалы *р≤2*). А және В сөз тіркестерін салыстыру кезінде олар сөз тізімдері түрінде ұсынылады, ал *әрбір, бұл, егер, осы, соған* сияқты қызметтік сөздер тізімдерге енгізілмейді. Сөздердің екі тізімі, егер олар сәйкес сөздердің барлық жұптарына тең болса, жуық эквивалентті болып саналады. Сонымен, *осы бір көйлек бояуы* сөз тіркесі { *бояуы, көйлек*} тізімі, ал *осы бір көйлектер бояуы* сөз тіркесі { *бояуы, көйлектер*} тізімі ретінде ұсынылады. Бірінші сөз жұбы үшін есептелген өңдеу қашықтығы (*бояуы* және *бояуы*) 0-ге тең, ал екінші сөз жұбы үшін (*көйлек* және *көйлектер*) 1-ге тең. Осылайша, сөз тіркестері *осы бір көйлек бояуы* және *осы бір көйлектер бояуы*, жуық эквивалентті болып саналады және мәтіндік нұсқалар ретінде анықталады.

Нұсқаларды автоматты түрде анықтаудың екінші тәсілінің бөлігі ретінде мәтіндегі терминдердің рұқсат етілген түрлендірілуі ресми түрде сипатталады, бұл әр түрлі нұсқаларды (атап айтқанда, сөздердің құрамымен және олардың арасындағы синтаксистік байланыстармен ерекшеленетін мәтіндік нұсқалар: *монохроматикалық толқын - синусоидалы толқын, қатты магниттік диск - қатты диск* және т.б.) толығырақ мен дәлірек анықтауға мүмкіндік береді.

Терминдерді түрлендіру ережелерін ресми сипаттау идеясы алғаш рет терминдер мен олардың нұсқаларын ағылшын және француз тілдеріндегі мәтіндер кешеніндегі нұсқаларды анықтау үшін құрылған FASTR жүйесінде жүзеге асырылды. Бұл жүйеде әр термин мәтінде оны тану ережесі ретінде сипатталады, ал терминологиялық нұсқалардың қалыптасу заңдылықтары жүйенің метаережелері арқылы бекітіледі. Мысалы, *water absorption* (су сіңімдігі) термині келесі ережемен сипатталады:

Rule: N1 N2 N3

<N1 lexicalization> = `N3´

<N2 lemma> = `water´

<N3 lemma> = `absorption´

Мұнда <N1 lexicalization>= `N3´құралымы терминнің негізгі сөзін *absorption* (N3) бекітеді және егер ол мәтінде кездессе, жүйе сипатталған *water absorption* (N2 N3) терминді анықтауға тырысуы керек екенін көрсетеді. Егер сіз ережені пайдаланып терминді тани алмасаңыз (мысалы, ...*water and sodium absorption*... сөйлемшесінде - *су мен натрийдің сіңуі*), содан кейін терминнің нұсқасын анықтау үшін метаережелердің көмегімен әрекет жасалады.

Метаережелер жаңа ережелер тудыруы мүмкін екендігімен сипатталады.

Мысалы, Метаереже

Metarule: Coor (Х1 Х2 ХЗ) = XI Х2 СЗ Х4 ХЗ

жалғаулық шылау арқылы бірнеше терминдердің қосылу жағдайларын сипаттайды және N2 N3 термин түріне N2 N3→ N2 SZ X4 N3 ережесін тудырады, егер бұл жүйенің сөздігінде мұндай талапқа жауап беретін термин болмаса, онда СЗ – жалғаулық шылау, ал X4 N3 құралымы жүйеге бұрыннан белгілі термин ретінде немесе үміткер-термині ретінде түсіндіріледі. Метаереже көмегімен...*water and sodium absorption*... сөйлемшесінде көрсетілген *water absoprtion* термині және *sodium absorption* үміткер-термині (мұнда СЗ -*and)* шығарылады

Осы тәріздес, FASTR жүйесі терминнің ішіне сөздерді енгізуге және қайта құруға байланысты терминдердің нұсқаларын анықтауға мүмкіндік беретін метаережелерді сипаттайды: *controlled delivery* – *controlled drug delivery* (*бақыланатын жеткізілім - бақыланатын дәрі жеткізілімі*), *book review* – *review of books (кітаби шолу - кітапқа шолу).* FASTR жүйесінде ережелер мен метаережелерді қолдану үшін мәтінді ішінара синтаксистік талдау негізінде жұмыс істейтін бағдарламалық модульдер жасалды [58].

Терминологиялық нұсқаларды қалыптастыру ережелерінің ресми сипаттамасы және оларды анықтау рәсімнен бөлу қолданыстағы қалыптар жиынтығын өзгерту арқылы терминдердің түрленуінің жаңа жағдайларын тану үшін МАӨ жүйелерін орнатуды жеңілдетеді. Бұл тәсілдің кемшіліктері әр нақты ТТ үшін ережелерді бейімдеу қажеттілігін қамтуы керек, өйткені олар тілге тәуелді болып келеді. Атап айтқанда, осы себепті FASTR жүйесін қазақ тіліндегі ҒТ мәтіндерге тікелей қолдану мүмкін емес.

# 1.4. Шығарылатын құралымдардың ресми сипаттамасы

Мәтіннен алынған тілдік құралымдардың ресми сипаттамасы белгілі бір артықшылықтарға ие болғандықтан, қазіргі заманғы МАӨ жүйелерін дамытуда қолданылатын ресімдеудің құралдарын қарастырған жөн.

Ең танымал және кеңінен қолданылатын аспаптық жүйелердің бірі - ТТ мәтіндерін автоматты түрде өңдеуге бағытталған GATE жүйесі шығарылған құралымдарды ресми сипаттау үшін арнайы Jape тілін пайдаланады. GATE ерекшелігі - өңделетін мәтіннің ішкі көрінісі: мәтіннің өзінен басқа (таңбалар тізбегі) жүйеде мәтін үзінділеріне аңдатпалар жиынтығы түрінде әр түрлі лингвистикалық ақпарат сақталуы мүмкін. Jape осы аңдатпаларды түрлендіру ережелерін жазуға қызмет етеді.

Jape ережелері келесідей: сол жақ → оң жақ. Сол жақта қажетті мәтіндік үзіндіні тану үшін тұрақты сөйлемдер жазылатын қалып көрсетіледі. Мысалы, мына қалып

{

мәтіндегі атаулардың бас әріптерін – екі элементтен тұратын (бас әріп және одан кейінгі нүкте) - *И.* түріндегі құралымды тануды анықтайды*,* ал келесі қалып

мәтіндегі атау септігінде келген сын есімдерді анықтауға қызмет етеді (пішінді жақшада жеке үзіндінің аңдатпасындағы шарттар келтірілген). Қалыптарда бұрын анықталған қалыптарды пайдалануға болады.

Jape-ережелерінің оң жағы ереженің сол жағымен танылған мәтін үзіндісінің аңдатпаларына қолданылатын түрлендірулерді сипаттайды. Мысалы, мәтіндегі *Иванов И*. есімнама түрінің аттарын тану және олардан фамилияларды алу ережесі келесідей:

Jape тілі жеткілікті икемді, бұл GATE жүйесін қолдана отырып, ТТ-мәтіндерден ақпарат алуға байланысты міндеттердің кең ауқымын шешуге мүмкіндік береді. Сонымен, GATE жүйесі негізінде бірнеше отандық жүйелер құрылды - атап айтқанда, OntosMiner [59] мәтіндерінен ақпарат алу үйірлік жүйелері және RCOFactExtractor ақпараттық-іздеу және аналитикалық жүйелерін әзірлеуге арналған кешенді құрал жасалды. Бұл жүйелер қазақ тіліндегі мәтіндерден деректі зат есімдерді: географиялық атаулар, есімнамалар, күн атаулары, мекенжайлар, заңды тұлғалардың есімдері және т.б., сондай-ақ олардың қатынастарын алуға бағытталған.

Танылатын мәтіндік бірліктерді сипаттаудың тағы бір ресми тілі отандық Alex [60] жүйесінде ұсынылды, сонымен қатар қазақ тіліндегі құжаттардың белгілі бір жанрларын (хаттар, хабарландырулар және т.б.) өңдеуге және олардан деректі зат есімдерді шығаруға бағытталған.

Alex жүйесінің қалыптарының негізгі элементтері ретінде танылады:

* мәтіндік сөз тұлғасы - көктем, азаматқа;
* оның бастапқы әріптерін анықтайтын сөздің символдық үлгісі - ауы... (мұнда үштік таңбалардың ерікті тізбегін білдіреді).

Жалпы алғанда, мұндай элементтер белгілі бір мәтіндік бірліктерді сипаттау және анықтау үшін қалыпқа біріктіріледі. Мысалы, мына қалып

[ауыл] - ауы...

ол келесі сөздерді сипаттайды: ауылдар, ауылдық. ауылнай және т. б., және келесі қалып

[Абай көшесі] = Абай көшесін

Абай көшесі және Абай көшесін секілді сөз тіркестерін сипаттайды.

Alex жүйесінің қалыптарында (GATE жүйесінің қалыптары сияқты) бұрын анықталған қалыптарды қолдануға болады. Мысалы, мына қалып

[орман алаңы] = орм... [алаң]

бірінші *орм* сөзінен басталып, келесі бөлігі [алаң] үлгісімен сипатталатын тіркестерді сипаттайды, мысалы, орман алаңы, ормандық алаң, орман алаңқайы.

Егер мәтінді өңдеу кезінде қалыппен сипатталған барлық мәтіндік үзіндіні емес, оның бір бөлігін ғана алу қажет болса, Alex жүйесінің тілінде контекстке тәуелді таңдау деп аталатын өлшемді орнатуға болады. Мысалы, келесі қалып жылдамдықтың сандық мәндерін таңдауға мүмкіндік береді:

[жылдамдық] = жылдамдық '[сан]' км/сағ

Мұнда [жылдамдық] - анықталған қалыптың атауы; жылдамдық пен км / сағ-бұл ТТ-дің нақты сөз тұлғалары, ал [сан] - санды белгілейтін бұрын анықталған қалыптың атауы. [Сан] элементі тырнақшаға алынғандықтан, ол қалыптан алынған нысан болып табылады, бірақ қалыптың қалған элементтері контекст ретінде қарастырылады.

Alex жүйесінің қалып тілінің айқын кемшілігі - оның танылатын мәтіндік бірліктердің тек лексикалық ерекшеліктерін және оларды қолдану құралымдарын сипаттауға ғана бағытталуы. Морфологиялық және синтаксистік ерекшеліктерді, мысалы, сөздердің қалыбына кіретін сөз таптарын белгілеу мүмкіндігі тілде жоқ. Одан өзге, Alex жүйесінің тілінде де, Jape тілінде де қалыптың әртүрлі элементтерінің грамматикалық сәйкестік байланысын нақты көрсетуге мүмкіндік беретін арнайы құралдар жоқ - бұл қазақ тілінің ғылыми-техникалық терминдері болып табылатын грамматикалық сәйкестік атаулы тіркестерді сипаттау үшін қажет.

Терминдерді алу әдістеріне жүргізілген шолудың нәтижелері бойынша терминдерді ҒТ мәтінінен алудың толықтығы мен дәлдігін арттыру үшін лингвистикалық және статистикалық критерийлердің белгілі бір тіркесімін қолдану қажет екендігі көрінеді.

Белгілі бір БҚ мәтіндер топтамасынан терминдерді алудың дәстүрлі міндетінен айырмашылығы, бұл диссертация жеке ҒТ мәтінінен әртүрлі терминдерді ғана емес, сонымен қатар олардың мәтіндегі барлық көріністерін алу міндетін қарастырады. Қарастырылып отырған мәтіннің көлемі бірнеше абзацтардан (статистикалық критерийлердің мүмкіндіктері айтарлықтай шектеулі болған кезде) бірнеше жүз бетке дейін өзгеруі мүмкін болғандықтан, мәселені шешуде лингвистикалық критерийлерді қолдануға баса назар аударылады.

МАӨ жүйесінде сөйлеудің ғылыми-техникалық стиліне тән терминдердің қолданылуы туралы лингвистикалық ақпаратты ұсыну үшін оны ресімдеу құралдары қажет. Диссертацияда осы мақсат үшін лексика-синтаксистік қалыптарды қолдану ұсынылады: олардың көмегімен терминдердің синтаксистік құрылымын, термин анықтамаларының құралымын және олардың синонимдерін енгізуді, сондай-ақ терминологиялық нұсқалардың қалыптасу ережелерін сипаттау ыңғайлы. Jape және Alex тілдерінің қалыптарынан айырмашылығы, лексика-синтаксистік қалыптар ресми грамматика теориясы мен лингвистикада қабылданған белгілерді қолданады, бұл оларды ықшамды әрі түсінікті етеді және терминологиялық ақпаратты автоматты түрде алу жүйелерін дамытатын мамандардың оларды құруы мен жетілдіруін жеңілдетеді.

Диссертацияда әзірленген терминдерді автоматты түрде алу рәсімдері және олардың талданатын мәтінге әр түрлі енуі алынған бірліктердің қалып түрінде ресми сипаттамасына сүйенеді. Бұрын айтылғандай, тілдік құралымдарды ресімдеу және оларды шығару рәсімдерінен бөлу бұл рәсімдерді басқа терминологиялық ақпаратты өңдеуге оңай орнатуға мүмкіндік береді.

Жалпы қабылданған терминдердің толықтығы мен дәлдігін және олардың берілген мәтінге енуін арттыру үшін рәсімдерде терминологиялық сөздік қолданылады және мұндай мәтіндерге тән терминологиялық нұсқалар ескеріледі. Өз кезегінде, мәтіннің жаңа терминдері үшін оларды алудың толықтығы мен дәлдігін арттыру терминдердің типтік құрылымын және анықтамалардың сипаттамалық құралымдарын ескере отырып қамтамасыз етіледі. Мәтінге жаңа терминдердің пайда болуын анықтаудың толықтығы мен дәлдігін арттыру терминдердің синонимдері мен терминологиялық нұсқаларын енгізу үшін ҒТ-мәтіндерге тән құралымдарды ескеру арқылы жүзеге асырылады. Сонымен қатар, жеке терминді қолдану жиілігі оның мәтінге барлық кіруін (оның синонимдері мен нұсқаларын қоса) ескере отырып есептеледі.

# **2 ТЕРМИНДЕРДІ ҚОЛДАНУДЫҢ ЛЕКСИКАЛЫҚ-СИНТАКСИСТІК ҚАЛЫПТАРЫ**

# 2.1 Ғылыми-техникалық терминдердің ерекшеліктері

Ғылыми-техникалық стильдегі мәтіндерге кіретін сөздер мен сөз тіркестерін шартты түрде жалпы ғылыми және арнайы деп бөлуге болады [61]. Жалпы ғылымиға барлық ана тілінде сөйлейтіндерге белгілі сөздер мен сөз тіркестері жатады. Олар ғылым мен техниканың әртүрлі салаларына ортақ ұғымдарды атайды (*ақпарат, мәселені шешу*), жеке мәтіндік үзінділерді байланыстыруға жауап береді (*демек, жоғарыда айтылғандарды қорытындылау*), пайымдауды рәсімдейді және реттейді (*айтарлық*, *не*, *осы себепті,*). Арнайы сөздерге мағынасы мен қолданылуы белгілі бір білім саласына бекітілген сөздер мен сөз тіркестері жатады: бірлік ұғымдарын білдіретін атаулар (*Ай, ASCII коды, Интернеттің инженерлік комиссиясы*) және осы білім саласының ұғымдарын атайтын терминдер (*митоз, индуктор, көрсеткіштердегі арифметикалық амалдар*) [62].

Терминдер бір сөзден (жеке сөздерден) және көп сөзден (сөз тіркестерінен) тұруы мүмкін. Бір сөзден тұратын терминдер көбінесе жалпы зат есімдер болып табылады (*компьютер, бағдарлама, толқын, гамма-сәулелер*) және сирегірек әр түрлі әріптік-символдық кешендер, соның ішінде әр түрлі қысқартулар (*БЭР, миниЭВМ, модем, р-п - ауысу, у-сәулелер*) түрінде кездеседі.

Көп сөзді терминдерге әдетте жалпы зат есімдер мен адъективтер (сын есімдер мен есімшелер) кіреді: *логикалық алгебра, ауырлық аномалиясы, автоматтандырылған оқыту жүйесі*. Кейде көп сөзді терминдер сан есімдерден (*термодинамиканың бірінші заңы, екінші тұтқырлық*), үстеу мен қосымшалардан (*қашықтық, төменнен жоғары бағдарламалау*), аббревиатуралар мен символдық кешендерден де (*бірліктер жүйесі* , *ішкі команда MS-DOS*) құралады. Көп сөзді терминдердің ұзындығы екіден бес-жеті сөзге дейін өзгереді; орташа ұзындығы бір-үш сөзден тұрады және қарастырылып отырған БҚ да байланысты.

Қазақ тіліндегі көп сөзді терминдердің арасындағы синтаксистік байланыстардың типтік түрлері - үйлестіру және басқару байланыстары. Қазақ тілінің терминдеріне тән басты белгі - зат есім мен оны анықтайтын адъектив арасындағы сәйкестік байланысы (*тұрақты тепе-теңдік, интегралды құрылғылар*), ал адъектив, әдетте, оған қатысты зат есімге, жіктелуі және септелуіне сәйкес келеді.

Зат есімдер арасындағы басқару байланысы айтарлықтай сирек кездеседі [63]. Қосымшасыз басқару байланысы болған жағдайда (*ақпарат алмасу*), басқа зат есімге анықтама ретінде әрекет ететін зат есім, әдетте, соңғысының оң жағында орналасады және көбінесе ілік септік тұлғасында болады (*материалдардың тозуы*). Басқаруды қосымшамен байланыстырған жағдайда (*бояғыш лазер*) оның оң жағында тұрған зат есімнің жағдайы қосымшаға байланысты: *кіші бағдарламаға жүгіну* – барыс септік, *санды бөлу* – табыс септік. Бірқатар терминдерде сәйкестік байланысы да, басқару байланысы да бар (*өзгермелі нүкте саны*).

ҒТ мәтінін өңдеуде қолданылатын компьютерлік терминологиялық сөздікте белгілі бір терминнің ұсынылғанына немесе ұсынылмағанына байланысты біз оны сәйкесінше сөздіктік немесе сөздіктік емес деп атаймыз.

Кез келген БҚ бойынша әрбір сөздік термині белгілі бір ұғымға бекітілген. Бұл тұжырымдаманың шекаралары көбінесе сөздікте берілген анықтама арқылы белгіленеді. Сөздіктік терминдер мағынасы қалыптасқан ұғымдарды атайды және олар үшін сөйлемнің қалыптасқан тілдік тұлға сипаттамасына ие. Егер сөздік термині мағынасы мен қолданылу саласы бойынша бірдей синонимді дублетке ие болса, онда әдетте екеуі де (термин де, оның дублеті де) терминологиялық сөздікте жазылады [64]. Дублеттердің мысалдары - *тіл білімі-тілтану-лингвистика* сөздері.

Сөздік емес терминді және оған сәйкес тұжырымдаманы байланыстыру, әдетте, (ол алғаш рет қолданылған) мәтіннің ішінде жүреді, сондықтан ҒТ мәтіндерінде терминдер нақты енгізілетін анықтамалық сөйлемшелер жиі кездеседі. Мысалы, мына сөйлемше

*Трансформациялық белгі - бұл ерекшеленетін басым белгі*... бұл *трансформациялық белгі* терминін анықтайды.

Сөздіктік емес терминдер тек тілдік сөйлем тұлғасында әлі қалыптаспаған дамып келе жатқан ұғымдарды атайды, сондықтан олар көптеген синонимдік атауларды қолданумен сипатталады (*белгі- таңба - тіркелген ынталандыру – шынайы ынталандыру – өзекті ынталандыру*). Терминдердің синонимдерін қалыптастыру ережелерін тұжырымдау немесе реттеу (соның ішінде бір ұғымның синонимдік атаулары әртүрлі ғылыми қауымдастықтарда пайда болатындықтан) іс жүзінде мүмкін емес, осыған байланысты мәтінде сөздіктік емес терминдердің синонимдері жиі енгізіледі. Мысалы, мына сөйлемшеде

*Бұл түзілімдерді гипержанрлар немесе гипержанр формалары деп атайық,* сөздіктік емес термин *гипержанр* деп енгізіледі және оның синонимі ретінде *гипержанр формасы* енгізіледі.

Оның ТТ мәтініндегі терминдерді қолданудың маңызды ерекшелігі (сөздіктік және сөздіктік емес) олардың нұсқаларын қалыптастыру мүмкіндігі болып табылады. Терминологиялық нұсқалардың ішінде біз мәтіндегі жеке терминді қолдану нұсқаларын ажыратамыз (*жиілік-контраст сипаттамасы - ЖКС, ток тасымалдаушылар - заряд тасымалдаушылар*) және бірнеше терминдерді қосу (2-ші және 3-ші деңгейдегі кэш — 2-ші деңгейдегі кэш және 3-ші деңгейдегі кэш).

ИмЕТ және физика салаларынан қазақ тіліндегі ҒТ мәтіндерінің (300-ге жуық мәтін) және терминологиялық сөздіктердің (15000-нан астам сөздік мақалалар) жинағын зерттеу [65], біз жүргізген қосымша, ғылыми-техникалық стильдегі мәтіндердің белгілі бір құралымдар мен терминдерді қолданудың сипаттамаларын растады. Осылайша, аталған лингвистикалық ақпаратты ресімдеу және оны ҒТ мәтіндерінен терминдерді автоматты түрде алу кезінде қолдану жеке терминдерді де, олардың өңделетін мәтінге әртүрлі енуін де танудың толықтығы мен дәлдігін жақсартуы керек.

# 2.2 Терминдердің құралымы мен қолданылу нұсқалары

Әдетте, ҒТ мәтіндерінде терминдер анықталатын немесе қолданылатын құралымдар (контексттер) үнемі қолданылады. Мысалы, мына түрдегі

*Мұндай реттілік мерзімдік қатар деп аталады*(2.1)

сөйлемшеде  *мерзімдік қатар* терминін енгізеді, ал мына

*Мақсатты енгізу дегеніміз - құру*,...(2.2)

сөйлемшесінде *мақсатты енгізу* терминін түсіндіреді. Осындай анықтамалық құралымдар аясында мәтінге енгізілген терминдер авторлық деп аталады.

Терминнің анықтамалық құралымында көбінесе терминнің өзі ғана емес, сонымен бірге ол үшін синоним де енгізіледі. Сонымен, мына сөйлемше

*Мұндай операцияларды тұжырымдамалық операциялар деп атаймыз*

*(тұжырымдамалық функциялармен) (2.3)*

авторлық термин ретінде *тұжырымдамалық операция* және оның синонимі ретінде *тұжырымдамалық функция* сөз тіркестерін енгізеді. Сондай-ақ, терминдердің синонимдері өз құралымдары бойынша да енгізілуі мүмкін. Мысалы, мына

*Гамма белгісін - еркін ынталандыру деп те атаймыз (2.3)* сөйлемшесінде *гамма белгісі* сөздіктік емес термин үшін *бос ынталандыру* синонимі енгізіледі, және келесі

... *арифметикалық-логикалық (атқарушы) құрылғы*...сөйлемшемінде *арифметикалық-логикалық құрылғы* сөздік термині үшін *атқарушы құрылғы* синонимі енгізіледі, мұнда терминнің өзі және оның синонимі үзілген түрде берілген, бұл оларды мәтіннен шығаруды қиындатады.

Біз зерттеген қазақ тіліндегі ҒТ мәтіндерінің жинағында авторлық терминдердің анықтамалары 55%-нда, ал синонимдердің нақты енгізілуі мәтіндердің шамамен 30%-нда кездесті; бұл ретте терминдердің анықтамаларының 100-ден астам әртүрлі құралымдары және олардың синонимдерін енгізудің 13 құралымы анықталды.

Мәтінде қолданылатын құралымдарды ресми түрде жазуға болады, мысалы, (2.2) және (2.3) сөйлемшелері схемалық түрде осылай жазуға болады

*T ретінде D-ны түсіндіреміз (2.4)*

*Т1-ді сонымен бірге Т2 деп атаймыз (2.5)*

Мұнда *ретінде, түсіндіріледі, жасаймыз, сонымен бірге, атау*- бекітілген сөз тұлғалары, Т және Т2 - келісілген атаулы топпен көрсетілген көмектес септік түрінде болатын негізгі сөз термин (немесе оның синонимі), Т1 - табыс септігіндегі атаулы топпен көрсетілген термин, D – атау септігіндегі атаулы топпен көрсетілген терминнің анықтамасы (түсіндірмесі). Мұндай қалыптық құралымдардың негізгі элементтері - бекітілген сөз тұлғалары (*ретінде, атау*) және термин мен оның анықтамасын білдіретін сөздер немесе сөз тіркестері.

Біздің зерттеуіміз сонымен қатар мәтінде қолданылған кезде жеке терминнің қиылғанын, қысқартылғанын және т.б., бір ұғымның әр түрлі сөйлем тұлғаларын құрайтындығын көрсетті - *2.1-суретті қараңыз*. Терминнің қалай өзгеретініне байланысты біз графикалық, орфографиялық, морфемалық, агглютинативтік, лексикалық-синтаксистік және қысқартылған нұсқаларды ажыратамыз.

|  |
| --- |
| *синоним (түбірлік морфеманы кесу)* |
| ҰҒЫМ  АЛҚ  АҚ  Арифметикалық құрылғы  арифметикалық-логикалық құрылғы  құрылғы  *синоним*  *(акроним)*  *Сөз тіркесін қию*  *синоним*  *(акроним)* |

Сурет 2.1. Терминнің мәтіндік нұсқалары

В сөз/сөз тіркесін А сөзінің/сөз тіркесінің графикалық нұсқасы деп атаймыз, егер В терминдері А сөздеріндегі әріптер регистрін өзгерту арқылы алынуы мүмкін болса, мысалы, *ағаш тамыры - АҒАШ ТАМЫРЫ*. Егер В-ны А құрамына кіретін сөздердің келісілген септелуі нәтижесінде А-дан алуға болатын болса: *жабынды, жабынмен - жабын* біз В-ны А - ның агглютинативті (жалғамалы) нұсқасы деп атаймыз. Графикалық және агглютинативтік нұсқаларды графематикалық және морфологиялық талдау негізінде анықтауға болады.

Егер В-ны оның сөздерінің біреуінің таңбалық құрамын өзгерту арқылы А-дан алуға болатын болса: тілдік тәуелді жүйе – тілге тәуелді жүйе, дастарқан – дастархан, біз В-ны А-ның орфографиялық нұсқасы деп атаймыз. Егер А сөздерінің біреуінің морфемалық құрамын өзгерту арқылы А-дан В алынуы мүмкін болса: орындалу - орындау, арнайы таңба – арнайы белгі, радио толқын жол – толқын жол, біз В-ны А-ның морфонұсқасы деп атаймыз. Біздің зерттеуімізге сәйкес [66] терминдердің орфографиялық нұсқасы және морфонұсқалары бір ҒТ-мәтінінде (терминдермен бірге) мүлде қолданбайды, тек сөздікте дублет ретінде жазылған сөздіктік терминдерінің жұптары түрінде ғана болуы мүмкін (мысалы, мультибағдарламалық есептеу жүйелері және монобағдарламалық есептеу жүйелері).

Егер А-ны құрайтын сөздердің лексикалық құрамы мен синтаксистік байланыстарының өзгеруі нәтижесінде А-дан В алынуы мүмкін болса: механикалық кернеу-кернеу, дисктік контроллер - диск контроллері, біз В-ны А-ның лексикалық-синтаксистік нұсқасы деп атаймыз. Мұндай нұсқаларды бір мәтінде қолдану ғылыми-техникалық прозаға тән және ол зерттелген топтама мәтіндерінің әрқайсысында байқалды.

Егер В-ны А құрамына кіретін сөздердің қысқаруына байланысты А-дан алуға болатын болса: деректер қорын басқару жүйесі — ДҚБЖ, емдік шара - емшара, біз В-ны А-ның қысқарту нұсқасы деп атаймыз. Қысқарту нұсқалары бір мәтінде де жиі кездеседі: сөздіктік емес терминдерге арналған қысқартулардың қалыптасуы және кейіннен қолданылуы зерттелген мәтіндердің жартысына жуығында кездеседі. Сөздіктік терминдерге келетін болсақ, олардың рұқсат етілген қысқартуларының басым көпшілігі терминологиялық сөздіктерге енгізілген.

Осылайша, ғылыми-техникалық мәтіндегі әртүрлі терминдерді қолдануды шығарудың толықтығын арттыру үшін, ең алдымен, сөздіктік және сөздіктік емес терминдерге тән лексикалық-синтаксистік нұсқаларды және сөздіктік емес терминдерге тән қысқарту нұсқаларын ескеру қажет.

Лексика-синтаксистік нұсқалардың басым көпшілігі белгілі бір ережелерге сәйкес құрылады, мысалы, терминнің бір немесе бірнеше сөзін қосу немесе тастау арқылы: *қолданбалы бағдарламашы - бағдарламашы, штаттан тыс жұмыс режимі - режим және* т. б. Мұндай ережелерді терминнің әрбір синтаксистік үлгісі үшін ресми түрде жазуға болады; мұндай қалыптық құралымдардың негізгі элементтері термин және оның нұсқасын білдіретін терминдерді білдіретін сөздер немесе сөз тіркестері болып табылады.

Қысқарту нұсқаларының ішінде, ҒТ мәтіндерінде бірнеше қысқарған сөздер кездесуі әдеттегі жағдай болады, сонымен қатар, олар әдетте мәтінде синонимдерді енгізу құралымдарына сәйкес және ұқсас түрде ресімделуі мүмкін тілдік құралымдар шеңберінде енгізіледі. Мысалы, мына сөйлемшеде

*...Күн сәулесіндегі жарылыстардан кейін планетааралық кеңістікте (ПК) байқалатын күн бөлшектерінің ағындары,...* осындағы *планетааралық кеңістік* термині үшін ПК қысқарту нұсқасын енгізеді.

Жеке терминнің нұсқаларынан басқа, ҒТ мәтіндерінде бірнеше терминдердің қосылыстары бар, олардың ішінде жалғаулық шылау көмегімен пайда болатын шылаулық қосылыстарды (және тыныс белгілерінің) және тек үйлестіру мен басқару байланыстарын қолдана отырып пайда болатын шылаусыз қосылыстарды бөліп көрсетуге болады. Қосылыстар зерттелген топтама мәтіндерінің басым көпшілігінде кездесті.

Шылаулық қосылыстар пайда болған кезде бірнеше терминдердің жалпы бөліктері қысқарады, бұл оларды мәтіннен автоматты түрде алуды қиындатады. Мысалы, мына сөйлемшеде *векторлық* немесе *растрлық дисплей* екі терминнен тұрады *векторлық дисплей* және *растрлық дисплей*, ал құралым болса *мекенжай, деректер және басқару құрсымы* ол үш термин негізінде жасалған - *мекенжай құрсымы, деректер құрсымы және құрсымны басқару*. Ұқсас жалғаулық құралымдарды бір (*және, не, немесе*) және қос шылаулар {*не-немесе, және-және, емес- және, екеуі де - және*) арқылы алуға болады.

Шылаусыз қосылыстар пайда болған кезде екі жағдай болуы мүмкін. Бірінші жағдайда бір-бірімен синтаксистік қатынасқа түсетін терминдер өзгермейді. Мысалы, екі терминден тұратын *материалдардың физика-химиялық қасиеттері* тіркесінде *физика-химиялық қасиеттері* мен *материалдары*, екі термин де ешқандай өзгеріссіз нақты бар. Екінші жағдайда қосылатын терминологиялық тіркестер өзгереді: ортақ бөліктерге байланысты үзіледі және/немесе қысқарады - 2.2-суретті қараңыз.

Мысалы, екі термин - *регистр* және *ішкі регистр* негізінде құрылған *ішкі регистрлердің разряды* тіркеуінде *регистрдің разряды* терминінің үзілуі және олардың жалпы бөлігі - *регистр* сөзінің бірігуі байқалады. Бұл жағдайда қосылысты құрайтын терминдерді алу да қиын.

*регистр*

*регистр*

*разрядтық*

*ішкі регистрлердің рарядтығы*

Сурет 2.2. Шылаусыз қосылыстың пайда болуы

Зерттелген ҒТ мәтіндерінде шылаулы қосылыстар жиі кездеседі. Кездескен қосылыстардың шамамен 87%-ы екі-үш терминнен тұрады, олардың құрылымы синтаксистік A N үлгісімен сипатталады (сын есім және зат есім) немесе NNgen (зат есім және ілік септіктегі зат есім).

Шылаусыз қосылыстарға келетін болсақ, негізінен оларды құрайтын терминологиялық тіркестер үзілмейді немесе қысқармайды. Терминдердің үзілуі және/немесе қысқаруы бар қосылыстар сирек кездеседі және әдетте сол A N және N Ngen синтаксистік үлгілермен сипатталған сөз тіркестерінен түзіледі (2.2 суретті қараңыз).

Терминдердің қарастырылған қосылыстары, сондай-ақ лексикалық-синтаксистік нұсқалар терминдердің белгілі бір ережелерге сәйкес ресми түрде жазылуы мүмкін әрбір синтаксистік үлгісі үшін жасалады.

Осылайша, қалыптық құралымдардың негізгі элементтері (терминдердің анықтамалары және олардың синонимдерін енгізу, сондай-ақ терминологиялық нұсқаларды құру ережелері) бекітілген сөз тұлғалары, сөздер мен сөз тіркестері болып табылады. Оларды ресми түрде жазу үшін сізге мыналар қажет:

* сөздің сөз таптарын белгілеңіз (зат есім, сын есім және т. б.);
* лексеманы оның бастапқы түлғасын көрсете отырып нақтылау;
* нақты сөз тұлғасын белгілеу;
* сөздің морфологиялық сипаттамаларының мағыналарын нақтылау (септелуі, саны, шағы және т. б.);
* сөздердің ретін белгілеу және олардың арасындағы синтаксистік байланыстарды, ең алдымен, сәйкес морфологиялық сипаттамалардың мәндерінің теңдігі ретінде көрсетілген синтаксистік сәйкестік байланысын көрсету.

# 2.3 Лексика-синтаксистік қалыптардың құрамы

Терминдердің құралымы мен қолданылу нұсқасын біз лексикалық-синтаксистік қалыптар түрінде рәсімдейміз [67]; қалып тілінің синтаксисін Бэкус-Наур қарапайым тұлғалар түрінде сипаттау А қосымшасында келтірілген, біз осы тілдің құралдарына жалпы сипаттама береміз.

Жалпы жағдайда лексика-синтаксистік қалып атау мен денеден тұрады. Атау дененің алдында теңдік белгісі арқылы жазылады; егер қалып басқа қалыптарда қолданылмаса, онда атауды көрсетпесе де болады. Қалып денесі сипатталған тілдік құралымды құрайтын элементтер тізбегін анықтайды және осы элементтердің синтаксистік сәйкестендіру шарттарын белгілейді. Мысалы, AN = A N қалыбының АN атауы мен A және N екі элементтен тұратын денесі бар.

Лексика-синтаксистік қалыптың негізгі элементтері – жол-элементі және сөз-элементі болады. Жол-элементі бекітілген сөз тұлғаларын, сондай - ақ қысқартулар мен тыныс белгілерін көрсетуге қызмет етеді, мысалы: «жол», «саны», « («.

Сөз -элемент сипатталған сөздің сөз табын орнатуға және қажет болған жағдайда оның морфологиялық сипаттамаларының мағыналарын көрсетуге мүмкіндік береді. Сөздің сөз табын көрсету үшін келесі белгілі символдық белгілер алынады: N - зат есім, V - етістік, А - сын есім, Pr –қосымша, Pn -есімдік және т. б. Сонымен, N сөз–элементі кездейсоқ зат есімді көрсетеді, ал егер бұрыштық жақшадағы сөз табын белгілегеннен кейін кейбір лексеманың бастапқы тұлғасы көрсетілсе, мысалы, N<файл>, содан кейін мұндай сөз–элемент зат есімнің барлық сөз тұлғаларын орнатады *файл: файл, файлды, файлға, файлдармен* және т. б.

Сөз-элементінде оның морфологиялық сипаттамалары нақтылануы мүмкін. Морфологиялық сипаттамаларды көрсету үшін символдық белгілер қолданылады: c-септік, t-уақыт, r-тұлға және т. б. мысалы, сөз–элемент A<c=ins, n = sing> кез келген сын есімді жекеше түрдегі тұлғаның көмектес септік түрінде, ал сөз–элемент \/< *түсінеді*, t=pres, P=3, m=ind>*түсіну* етістігін осы шақ, ашық рай, 3 жақ түрінде сипаттайды, яғни сөз тұлғалары: *түсінеді* немесе *түсініледі*.

Қалып элементтерінің реті сипатталған құралымдағы сөздердің ретіне сәйкес келеді: мысалы, сын есім мен зат есімнен тұратын терминнің жалпы синтаксистік үлгісі (*уақыттық қатар*) А N қалыбымен сипатталады.

Үлгі элементтері арасындағы келісудің синтаксистік байланысын орнату үшін келісу шарттары қолданылады. Шарттар барлық элементтерді сипаттағаннан кейін бұрыштық жақшадағы сәйкес морфологиялық сипаттамалардың мәндерінің теңдігі ретінде жазылады, мысалы:А N<А.с=N.с,А.n=N.n,А.g=N.д>

Бұл жазба А сын есімі мен N зат есімі септелуі, жағы және саны келісілуі керек дегенді білдіреді. Мұндай келісімді қысқаша сипаттауға болады: АN <А=N>

Сипатталған құралым бір сөз табынан бірнеше түрлі сөздері болуы мүмкін болғандықтан, оларды ажырату үшін сандық индекстер қолданылады. Мысалы, *мақсаттың енгізілуі* термин түрі қалыппен сипатталуы мүмкін N1 N2<с=gen> ,ол екі түрлі N1 және N2 зат есімдерін қамтиды, олардың екіншісі ілік септікте (бұл тіркесте болатын басқару байланысы нақты көрсетілмеген) болады.

Сипатталған тілдік құралымның бірнеше мүмкін нұсқаларын көрсету үшін қалыптың балама элементі қолданылады, нұсқалар | белгісі арқылы көрсетіледі. Мысалы, Т деп аталатын қалып Т=N1| А1 N1 <А1=N1>| N1 N2<с=gen>терминнің үш мүмкін синтаксистік үлгісін сипаттайды: N1, А1N1<А1=N1> және N1 N2<с=gen>.

Опциондық элемент қалыптың кейбір элементінің ресімделетін құралымға енуінің міндетті еместігін көрсету үшін қолданылады. Мысалы, D деп аталатын қалыпта D = [Pn < осындай>] N <Pn=N>мұнда *осындай* есімдігінің опционалдық (міндетті емес) пайда болуы көрсетілген. Бұл қалып, атап айтқанда, мысалдардан (2.1) және (2.2) терминдердің анықтамаларын қанағаттандырады: *осындай дәйектілік* және *құру.*

Сондай-ақ, қалыптың элементі басқа қалыптың данасы (атауы) болуы мүмкін. Жоғарыда сипатталған қалыптарды Т және Б атауларымен қолдана отырып, түрдің құралымы (2.4) лексикалық-синтаксистік қалып ретінде жазылады:

«атайық» Т1<с=ins>V1<түсінемін, t=pres, р=1, m=ind>D1<с=nom>

<Т1.n=V1.n>,

мұндағы «атайық» - жол-элемент, V1<түсінемін, t=pres, r=1m=ind> - элемент-екі сөз формасын сипаттайтын сөз: *түсінемін* және *түсінеміз*, Т1 және d1 - атаулары бар Т және D қалыптардың даналары сәйкесінше,<Т1.n=V1.n> - T1 терминінің және VI етістіктің грамматикалық санын үйлестіру шарттары. Синонимді енгізу құралым (2.5) қалып ретінде келесідей жазылады:

Т1<с=асс> «боламыз» «деп» «түсіну» Т2<с=ins>

Бұл қалыптарда Т және D атаулары бар қалыптардың даналық элементтері үшін септіктің нақтылануы көрсетілгенін ескеріңіз: T1<с=асс>, D1<с=nom>. Бұл мүмкін болу үшін қалыптарды сипаттау кезінде Т және D септіктері параметр ретінде көрсетілуі керек. Параметрлер анықталған қалыптың барлық элементтерінен кейін жақшаға жазылады және оның элементтерінің морфологиялық сипаттамаларын кейінірек нақтылау немесе сәйкестендіру жағдайында параметрлерді қолдану үшін қажет. Мысалы, егер D қалыбының параметрлері N сөз-элементінің барлық морфологиялық сипаттамалары болса:

D= [Pn<деп>] N <Pn=N> (N),

онда элемент D<с=nom> атау септігіндегі есімдік пен зат есімнің келісілген тобын сипаттайды.

Лексика-синтаксистік қалыптар тілінің сипатталған құралдарын біз ресми жазба үшін қолдандық:

* терминологиялық сөздер мен сөз тіркестерінің негізгі синтаксистік үлгілері;
* ғылыми-техникалық прозаға тән құралымдар, терминдердің анықтамалары және олардың синонимдерін енгізу, сондай-ақ олардан терминдерді алу ережелері;
* сөздіктік терминологиялық сөздер мен сөз тіркестері;
* терминдердің лексикалық-синтаксистік нұсқаларын қалыптастыру ережелері;
* терминдер мен олардан терминологиялық сөздер мен сөз тіркестерін алу ережелерінің қосылыстарының типтік синтаксистік үлгілері.

Алынған үлгілердің мысалдарын қарастырыңыз.

# 2.4 Құралым және терминдер нұсқаларының қалыптары

Жоғарыда аталған тілдік құралымдардың лексикалық-синтаксистік қалыптарын сипаттау үшін біз терминдердің жеті жиілік синтаксистік үлгілерімен шектелдік. Үлгілер 2.1-кестенің бірінші бағанында ұсынылған, онда АР-адъектив, яғни толық тұлғалы сын есім немесе есімше. Кестеде сонымен қатар зерттелген мәтіндер топтамасында және терминологиялық сөздіктерде қарастырылып отырған үлгілерге сәйкес терминдерді қолдану статистикасы жинақталған[68]. Кестенің екінші және үшінші бағандарында сәйкесінше мәтіндер мен сөздіктердегі олардың жалпы санынан әр үлгіні қанағаттандыратын терминдердің үлесі көрсетілген. Қалған 23,2% (мәтіндер үшін) және 16,2% (сөздіктер үшін) терминологиялық тіркестер 100-ден астам басқа синтаксистік үлгілермен сипатталады.

Кесте 2.1. Терминдердің негізгі синтаксистік үлгілері

|  |  |  |  |
| --- | --- | --- | --- |
| Үлгі | Мәтіндер | Сөздіктер | Мысалдар |
| АР N | 28,5 % | 35,9 % | Виртуалдық панель; альбедті мюон |
| N | 17,0% | 23,3 % | аутентификация; дыбыс өткігіштік |
| N Ngen | 10,8 % | 14,1 % | Алдын алу кезеңі; шуыл генератор |
| N APgenNgen | 4,5 % | 3,3 % | геодезиялық сызықтар ауытқуы |
| AP N Ngen | 7,2 % | 2,8 % | Серфотронды жеделдету механизмі |
| AP APN | 6,5 % | 2,6 % | Қатты магниттік диск |
| N NgenNgen | 2,3 % | 1,8% | Деректердің тұтастығын бұзу |
| 76,8 % 83,8 % | | | |

SP - терминдердің 7 жиілік синтаксистік үлгісін сипаттайтын лексикалық-синтаксистік үлгінің атауы болсын:

SP = N1 (N1) | AP1N1 <AP1=N1>(N1) |NlN2<c=gen>(N1) |

AP1 AP2 N1 <AP1=AP2=N1>(N1) |

AP1N1 N2<c=gen><AP1=N1>(N1) |

N1 AP2 N2<c=gen><AP2=N2>(N1) |

N1 N2<c=gen> N3<c=gen>(N1) ,

мұндағы АР = А (А) | Ра (Ра) қалыбы адъектив ұғымын анықтайды. Қажет болса, SP қалыбы белгілі бір БҚ терминологиясының ерекшеліктерін, атап айтқанда, келесі үлгіні ескере отырып, басқа синтаксистік термин үлгілерімен оңай толықтыруға болатындығын ескеріңіз

N1 «с» АР2 N2<с=ins><АР2=N2>

(*дүкен жадысы бар автомат* термин түрі үшін) немесе үлгі

N1 N2<с=gen>N3<с=gen> «және» N4<с=gen>

( *масса және энергия өзара байланысы заңы* термин түрі үшін).

Ғылыми-техникалық прозаға тән құралымдарды ресімдеу авторлық терминдердің анықтамаларының 64 қалыбын (олар 100-ге жуық әртүрлі құралымды қамтиды) және терминдердің синонимдерін енгізудің 7 қалыбын (13 құралымды сипаттайды) берді. 2.2-кестеде алынған лексикалық-синтаксистік қалыптардың мысалдары келтірілген; үстеме мысалдар Б қосымшасында келтірілген.Қалыптар мен мысалдардың үзінділерінде авторлық терминдердің асты сызылады.

Кесте 2.2. Тілдік құралым қалыптарының мысалдары

|  |  |  |
| --- | --- | --- |
| *№* | Қалып | Құралым мысалдары |
| *1* | *Terml<c: =ins>«әдетте»*  *«атау» Defin1<c=acc> #*  *Terml* | *Жұқа клиент әдетте желінің компьютер- клиентін атайды...* |
| *2* | *Definl<c=nom>V1<аталу, t=pres, p=3, m=ind>Terml<c=ins>*  *# Terml* | *Мұндай реттілік уақытша қатар деп аталады* |
| *3* | *«деп» Terml<c=ins>*  *V1<ұғыну, t=pres, p=3, m=ind>Definl<c=nom># Terml* | *Пайдаланушының жұмыс үстелі деп механизмді ұғынамыз...* |
| *4* | *Term1<c=nom> [«-»] «бұл»*  *Defin1<c-nom># Term1* | *Ақпараттық ресурс - бұл мәтіндік файлдар жинағы* |
| *5* | *AP1» (« [«немесе»] AP2») » SP1 <AP1=AP2=SP1>*  *# API SP1 <AP1=SP1>,*  *AP2 SP1 <AP2=SP1>* | *...физикалық (метрикалық) кеңістіктегі "қашықтықты" анықтайды* |
| *6* | *SP1» («SP2») »<SP1.c=SP2.c>*  *# SP1, SP2* | *...Ақпараттық жүйелерді (АЖ) құру.* |

Терминдердің анықтамалық қалыптарына Term және Defin атаулары бар қалып даналары кіреді.Олардың біріншісі анықталған терминнің синтаксистік құрылымын бекітеді

SP1 [[«,»] «немесе» [«қарапайым»] SP2] <SP1.c=SP2.с> (SP1)

#SP1, SP2 ,

ал Defin деп аталатын үлгі құралымның терминді анықтайтын құрылымын белгілейді. Ол бірнеше баламаларды қамтиды, атап айтқанда:

[Pn<кез келген>|Pn<осыындай>|Pn<бұл>] SP <Pn=SP> (SP) ,

яғни, анықтама ретінде SP қалыбымен сипатталатын атаулы тіркес болуы мүмкін, оның алдында *әрбір*, *осындай* немесе *бұл* деген есімдік болуы мүмкін. Term және Defin қалыптарының толық сипаттамасы Б қосымшасында берілген.

Терминдердің танылған тілдік құралымдардан алу ережелері (және қажет болған жағдайда олардың анықтамалары) қалып тілінде болады, # белгісінен кейін қалып денесінде жазылатын алу қалыптар арқылы беріледі.

Мысалы, мына қалыпта

AP1»(« [«немесе»] АР2 «)» SP1<AP1=AP2=SP1>

# AP1 SP1 <AP1 = SP1>, AP2 SP1 <AP2=SP1>

AP1SP1 алу қалыптары <AP1=SP1 > және AP2 SP1 <AP2=SP1> қалып арқылы танылған AP1, SP1 және A2 элементтерінің құралымын таңдауды және олардан екі AP1SP1 және A2 SP1 сөз тіркестерін құрауды енгізеді. Жоғарыдағы қалыпты қолдану арқылы келесі сөйлемшеден

...*физикалық (метрикалық) кеңістіктегі» қашықтықты « анықтайды* осыдан екі термин бөлінеді: *физикалық кеңістік* және *метрикалық кеңістік*. Осылайша, лексика-синтаксистік қалыптардың құралдарын қарастыру мәтіннен терминдерді және оларды үзілген жағдайда да алуға мүмкіндік береді.

Лексика-синтаксистік қалыптар сөздіктік терминдерді енгізудің ыңғайлы құралы болып шықты. Қаалыптар арқылы ИмЕТ және физика сөздіктеріндегі терминдер сипатталды [69]. Сөздіктік терминдер қалыптарының мысалдары 2.3-кестеде келтірілген.

Кесте 2.3. Сөздіктік терминдері қалыптарының мысалдары

|  |  |
| --- | --- |
| Қалып | Терминдер |
| *N1<масс-спектр>* | масс-спектр |
| *N1<алгебра>N2<логика,с=gen>* | логикалық алгебра |
| *А1<жартлай өткізгіш> Ра1<есте сақтау>N1<құрылғы><А1=Ра1=N1>* | Жартылай өткізгішті сақтау құрылғысы |
| *N1<мекенжай> {N2<топ, с=gen> | N2<қайтару,с=gen> | N2<құрылғы,с=gen>}* | Команданың мекенжайы, қайтару мекенжайы, құрылғының мекенжайы |
| *N1<критерий>*  *[«Лоусона»| N2<ұқсастық,с=gen> |*  *«Рэлея»]* | критерий, Лоусон критерийі, ұқсастық критерийі, Рэлей критерийі |

Сөздіктік терминдер қалптарының басты ерекшелігі-бір қалып бір сөзден немесе сөз тіркесінен басталатын бірнеше терминдерді сипаттай алады. Мысалы, мына қалыпта

N1<драйвер>

[ N2< пайдаланушы,с=gen> | N2<терминал,с=gen> ]

*драйвер* сөзі кедесетін терминдерді сипаттайды: драйвер, пайдаланушы драйвері және терминал драйвері, ал қалып

A1 < Мәңгілік > N1 < қозғалтқыш>

{Num1 < екінші> / num1 < бірінші>}

N2 < тегі, с=gen><A1=N1, Num1=N2>

*мәңгілік қозғалтқыш* сөз тіркесі кездесетін терминдерді сипаттайды: *екінші тектегі мәңгілік қозғалтқыш* және *бірінші тектегі мәңгілік қозғалтқыш*. Соңғы қалып жазбасында міндетті элемент деп аталатын бар (пішінді жақшаларға алынады), олардың біреуі сипатталған құралымда пайда болуы керек баламаларды топтастырады.

Сөздіктерден алынған сөздіктік қалыптар жиынтығында 4321 ИмЕТ терминін сипаттайтын 1725 қалып және физика бойынша 3444 терминді сипаттайтын 2010 қалып бар. Сөздіктік терминдер қалыптарының үстеме мысалдары В қосымшасында келтірілген.

Лексикалық-синтаксистік қалыптар түріндегі лексикалық-синтаксистік сөздіктік терминдер үшін олардың дублеттері де сипатталған. Мысалы, физика терминдері үшін *толық өткізгіштік - адмитанс* жазылған қалып мынадай

A1 < толық > N1 < өткізгіштік><A1=N1 > # N2 < admitance> , мұнда қалып денесі терминді белгілейді, ал # таңбасынан кейін оның шығару қалыбы ретінде дублеті жазылады. Қазіргі уақытта сөздіктер негізінде ИмЕТ және физика салаларынан сәйкесінше 1006 және 394 жұп <термин-дублет> алынды. В қосымшасында < термин-дублет> жұптық қалыптардың үстеме мысалдары келтірілген.

Лексикалық-синтаксистік нұсқалардың қалыптасу ережелерін сипаттайтын қалыптар келесідей ұйымдастырылған: әр қалып денесі терминдердің типтік синтаксистік үлгілерінің бірін, ал алу қалыбы оның болжалды мәтіндік нұсқаларының синтаксистік үлгілерін көрсетеді. N1 N2<C=gen > түрінің терминдеріне сәйкес қалыпты қарастырамыз (мәтіндік нұсқалар терминнің синтаксистік үлгісінен кейін, жолдар бойынша жазылады):

N1 N2<с=gen> #

N1,

N1 N4<с=gen><Syn (N2,N4)>,

N3 N2<с=gen><Syn(N1,N3)>,

А1 N1 <Der(А1,N2),А1=N1>

Мәтіндік нұсқасы бар әрбір жол лексика-синтаксистік нұсқалардың көпшілігі жасалатын негізгі ережелердің бірін білдіреді.

Біріншіден, нұсқаны сөзді немесе сөз тіркесін қосу немесе тастау арқылы бастапқы терминнен алуға болады: N1 N2<c=gen> термин түрі үшін N1 (*кристалл симметриясы* - *симметрия, деректерді енгізу - енгізу*) нұсқа болып табылады.

Екіншіден, нұсқаны бір сөзді/сөз тіркесін синонимге ауыстыру арқылы терминнен алуға болады (бұл бастапқы термин сияқты синтаксистік құрылымы бар, бірақ басқа лексикалық құрамы бар сөз тіркестерін шығарады). N1 N2<c=gen> үлгісі бар термин үшін N1 N4<c=gen>(*деректерді беру - ақпаратты беру*) немесе N2<c=gen> (*мекенжай белгісі - мекенжай маркері*) нұсқа болады, мұндағы N4 сөзі N2 үшін синоним, ал N3 сөзі N1 үшін синоним болып табылады. Мұндай шарттарды тексеру сөздіктік шарттарын қолдана отырып қалыпқа жазылатын сөздікке жүгінуді қажет етеді: Syn(N2, N4), Syn(N1, N3).

Үшіншіден, нұсқаны сөз тіркесінің синтаксистік байланыстарын бір уақытта өзгерте отырып, сөзді басқа сөз табының бір түбірімен ауыстыру арқылы терминнен алуға болады (бұл жағдайда терминнің синтаксистік құрылымы мен лексикалық құрамының өзгеруі байқалады). N1 N2<c=gen>түрінің терминдері үшін A1N1 <A1=N1> осындай нұсқа бола алады (*ығысу толқындары - ығысқан толқындар, шу генераторлары - шуыл генераторлары*). Сонымен қатар, A1 және N2 бір түбірлі туынды сөздер - Der (A1, N2) түріндегі сөздік шарттары арқылы қайтадан жазылатын дериваттар болып табылады.

Синонимдер мен дериваттар <термин-дублет> жұптарына ұқсас қалыптармен сипатталғанын ескереміз. Сөздіктер негізінде ИмЕТ саласынан 458 жұп синонимдер және физика саласынан 392 жұп синонимдер, сондай-ақ сәйкесінше 237 және 144 жұп дериваттар жинақталған. Синонимдер мен дериваттар сөздіктік қалыптарының мысалдары В қосымшасында, ал біз ескеретін терминдердің синтаксистік үлгілеріне арналған лексикалық-синтаксистік нұсқаларды құрудың барлық ережелері Г қосымшасында келтірілген.

Терминдердің лексикалық-синтаксистік нұсқаларын қалыптастыру ережелерінің қарастырылған қалыптарына ұқсас, бірнеше терминологиялық тіркестердің қосылу қалыптарыі және олардың құрамдас терминдерін алу ережелері сипатталған. Мұндай қалыптың денесі қосылыстың синтаксистік үлгісі болып табылады және осы қосылысты құрайтын терминдер (олардың синтаксистік үлгілері) алу қалыптарымен берілген.

Мысалы, мына қалып

А1 А2 N1 <А1=А2=N1>

# А1 N1 <А1=N1>, А2 N2 <А2=N2>

екі A1 N1 және A2 N2 түрлеріндегі терминнің шылаусыз қосылыстарын сипаттайды. Бұл қалып мына сөйлемшеден: *адрестік микропроцессорлық жад* келесі терминдерді - *адрестік жад* және *микропроцессорлық жад* алуға мүмкіндік береді

Шылаусыз қосылысты сипаттайтын қалыптың мысалы ретінде мына қалыпты қарастырыңыз

«қалай» N1 «,» «солай» «және» N2 N3<с=gen> #

N1 N3<с=gen>,

N2 N3<с=gen> ,

Бұл N Ngen түрінің екі терминін *қалай – солай және* қос шылауы арқылы біріктіру жағдайын сипаттайды. Көрсетілген қалып негізінде *бағдарламаны қалай әзірлеу және солай қолдау* сөйлемшесінен екі термин бөлінуі мүмкін: *бағдарламаны әзірлеу* және *бағдарламаны сүйемелдеу*. Көп сөзді терминдердің типтік қосылыстарының лексикалық-синтаксистік қалыптары Г қосымшасында келтірілген.

Осылайша құрылған лексикалық-синтаксистік қалыптар жиынтығы лингвистикалық ақпараттың негізін құрайды, оған осы диссертацияда жасалған терминдердің әр түрлі қолданылуын қазақ тіліндегі ҒТ мәтінінен алу рәсімдеріне сүйенеді. Рәсімдердің жұмысы негізінен өңделетін мәтінге лексикалық-синтаксистік қалыптарды қабаттастыруға, ондағы қалыптармен сипатталған мәтіндік үзінділерді іздеуге және осы үзінділерден тиісті терминдерді қолдануға дейін азаяды.

# **3 ТЕРМИНДЕРДІ ТҰТЫНУДЫ АЛУДЫҢ РӘСІМІ**

# 3.1 Рәсімдердің жұмыс схемасы

Лексикалық-синтаксистік қалыптар түрінде танылатын терминдер, сондай-ақ құралымдар мен оларды қолдану нұсқалары туралы әр түрлі лингвистикалық ақпарат рәсімделгендіктен, диссертацияда алу рәсімдері жасалды, олардың әрқайсысы терминді тұтынудың белгілі бір түріне сәйкес келеді. Рәсімдер 3 топты құрайды:

1) терминдер кіруінің алу рәсімдері:

getDictTerms - сөздіктік терминдерді алу;

getNonDictTerms - сөздік емес терминдерді алу.

2) Терминдерді кейін алу үшін, тілдік құралымдарды тану рәсімдері:

getAuthTerms - авторлық терминдерді алу;

getSynTerms - терминдердің синонимдерін алу;

getTermsfromCombs - қосылыстардан терминдерді алу.

3) Терминдердің мәтіндік нұсқаларын тану рәсімдері: getVarsforTerms.

getDictTerms рәсімі сөздіктік терминдер қалыптарымен жұмыс істейді, ал getNonDictTerms рәсімі - терминдердің типтік синтаксистік үлгілерін сипаттайтын қалыппен жұмыс істейді. Қалыптарды мәтінге қабаттастыру нәтижесінде мәтінге танылған терминдердің кіруіне сәйкес келетін мәтін үзінділер алынады.

getAuthTerms, getSynTerms және getTermsfromCombs рәсімдері кіру үшін денелері танылатын тілдік құралымдарды сипаттайтын қалыптарды алады, ал алу қалыптары осы құралымдардан терминдерді алу ережелерін белгілейді. Бұл жағдайда шаблон денесінің мәтініне қабаттасу танылған құралымдардың мәтініне кіруді білдіретін мәтін үзінділерін береді, содан кейін терминдердің өзі алынады.

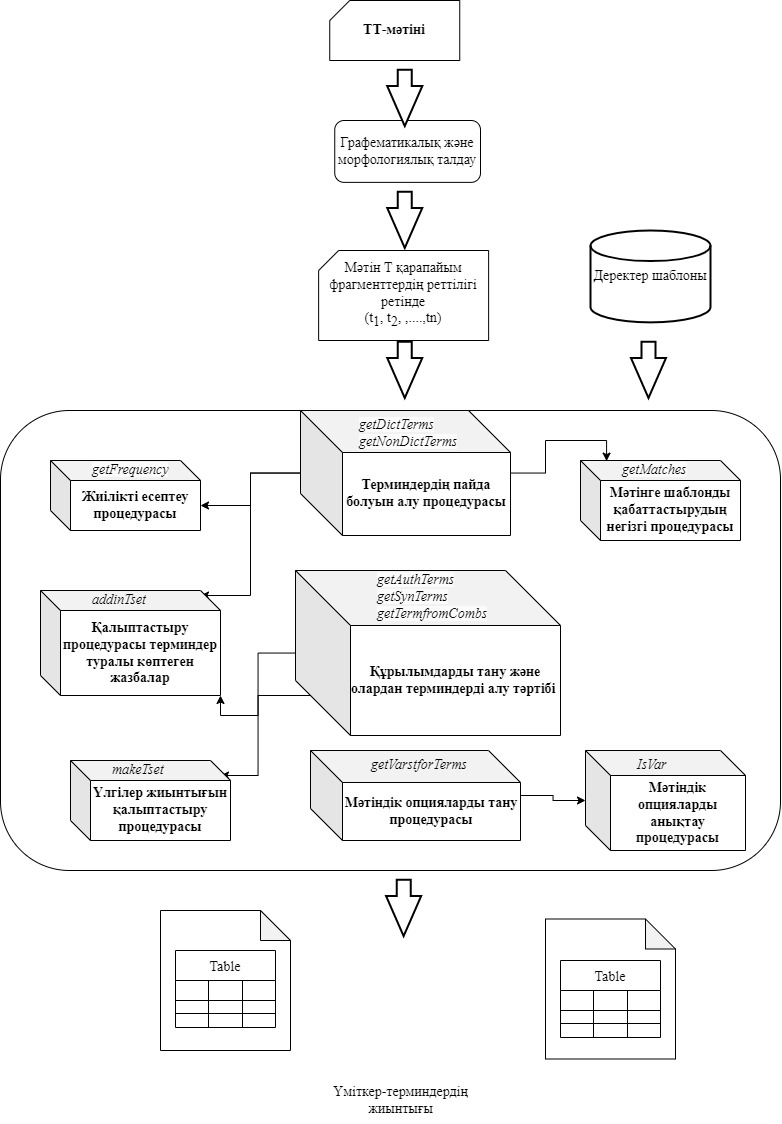
getVarsforTerms рәсімі денелері терминдердің синтаксистік үлгілерін сипаттайтын қалыптармен жұмыс істейді, ал алу қалыптары ықтимал мәтіндік нұсқалардың синтаксистік үлгілерін орнатады. Мұнда қалып денесінің қабаттасуы терминдерге сәйкес келетін мәтіндік үзінділерді алуға мүмкіндік береді, содан кейін олардың мәтіндік нұсқалары жасалады.

Жоғарыда аталған рәсімдердің әрқайсысының нәтижесі үміткер-терминдердің жиынтығы болып табылатындығын және бұл жиынтықтар бір-біріне сәйкес келуі мүмкін екенін ескереміз, өйткені бір сөз/сөз тіркесі кейде әртүрлі рәсімдермен анықталады. Мысалы, сөздіктік емес терминдерді алу рәсімі арқылы алынған сөз тіркесін терминдерді қосылыстардан алу рәсімі арқылы кейбір қосылыстардың бөлігі ретінде де анықтауға болады.

Мәтінге графематикалық және морфологиялық талдау жасалды деген болжаммен терминологиялық ақпаратты алу рәсімдерін сипаттайтын боламыз, ал графематикалық талдау барысында барлық әріптер бір регистрге келтіріліп, мәтіннен бос орын және басқару таңбалары (бос орын, қойынды, жаңа абзац белгісі және т.б.) жойылады. Мәтінді өңдеудің жалпы схемасы 3.1-суретте көрсетілген.

Сипатталған терминдерді қолдану рәсімдері мәтіннің үстірт талдауын жүзеге асырады және өз жұмысында getMatches негізгі рәсіміне (мәтінге қалып қоюға жауапты) және келесі көмекші рәсімдерге сүйенеді:

* getFrequency - терминологиялық тіркестердің бір-біріне енгізілуін ескере отырып, мәтіндегі терминдерді тұтыну жиілігін есептеу;
* addinTSet - алынған терминдер туралы жазба жиынтығын қалыптастыру;
* makePSet - (мәтінге қабаттастыру үшін) қалыптардың қосымша жиынтығын құру;
* isVar термині - терминдердің мәтіндік нұсқаларын анықтау.



Сурет 3.1. Мәтінді өңдеудің жалпы схемасы

Графематикалық және морфологиялық талдау жүргізілгеннен кейін талданатын Т мәтін - қарапайым мәтін үзінділерінің тізбегі ti-сөз тұлғалары және оларды бөлетін таңбалар: Т=(t1, t2 , … , tnT ) бастапқы мәтінмен бірдей тәртіпте орналасқан. Сөз тұлғасы болып табылатын әрбір қарапайым ti үзіндісіне бұл сөздің сөз табы, бастапқы тұлғасы және оның *xi* морфологиялық сипаттамаларының мәндерінің жинағы (тегі, саны, шағы және т.б.) белгілі. Мысалы, сезбек сөз тұлғасының үзіндісіне морфологиялық талдау барысында бұл зат есім, оның бастапқы формасы сез, оған –бек қосымшасы жалғанған, ал морфологиялық сипаттамалары келесі мағыналарға ие екендігі анықталады: тұлғасы – туынды, саны - жекеше, қатыстығы - жансыз, септігі – атау септік немесе табыс септік (берілген сөз үшін қалған морфологиялық сипаттамалардың мәні жоқ).

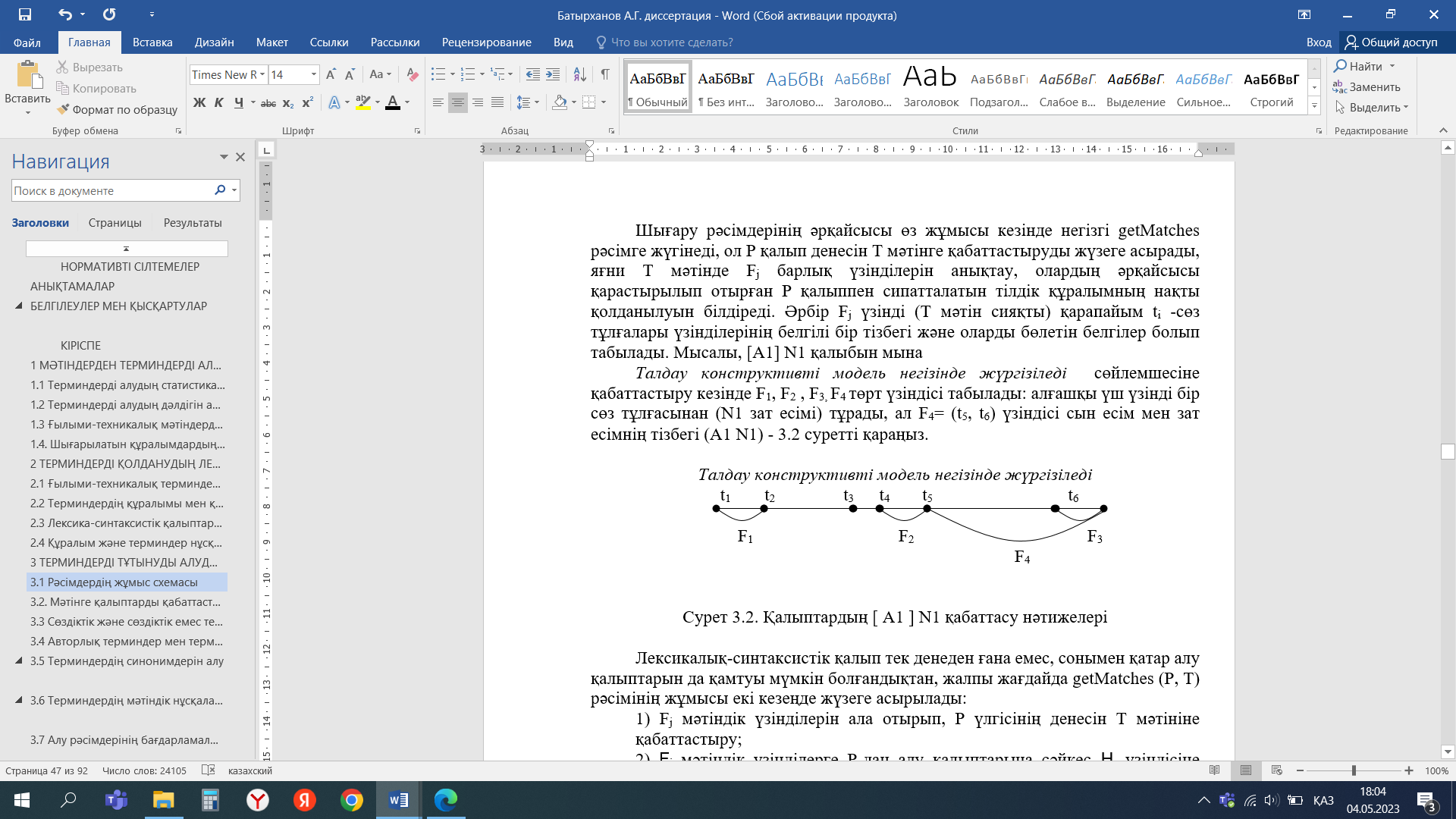
Сипатталған терминдерді тұтынуды алу рәсімдерінің әрқайсысының кірісіне (getVarsforTerms рәсімінен басқа) Т=(t1, t2 , … , tnT ) талданатын мәтін және S={P1, P2 , … , PnS } қалыптардың жиынтығы келеді. Кез келген рәсімнің міндеті S жинағындағы әрбір Рi қалыбын T мәтініне қабаттастыру және осы процесте алынған нәтижелерді кейіннен өңдеу; рәсімнің шығуында - MR={mr1, mr2,..., mrnM} жиынтығы олардың әрқайсысын тұтыну жиілігі туралы ақпарат бар алынған үміткер-терминдер туралы жазбалар - 3.1-кестені қараңыз.

Кесте 3.1. Терминологиялық ақпаратты алу рәсімдері

|  |  |  |
| --- | --- | --- |
| getDictTerms – сөздіктік терминдерді алу | | |
| Мәтін және сөздіктік терминдерінің қалыптары | мәтіннің сөздіктік терминдері оларды тұтыну жиілігімен | |
| getNonDictTerms - сөздіктік емес терминдерді алу | | |
| терминдердің негізгі синтаксистік үлгілерін сипаттайтын мәтін және қалып | сөздіктік емес терминдер оларды тұтыну жиілігімен | |
| Кіру | Шығу | |
| getAuthTerms – авторлық терминдерді алу | | |
| Мәтін және терминдерді анықтау қалыптары | | авторлық терминдер оларды тұтыну жиілігімен |
| getSynTerms- синонимдерді алу | | |
| Мәтін және синонимдерді енгізу қалыптары | | терминдер және олардың синонимдері оларды тұтыну жиілігімен |
| getTermsfromCombs – қосылыстардан терминдерді алу | | |
| Мәтін және термин қосылыстарының қалыптары | | терминдер (қосылыстардан) оларды тұтыну жиілігімен |
| getVarsforTerms - мәтіндік нұсқаларды тану | | |
| Үмітер-терминдер және мәтіндік нұсқалардың қалыптасу ережелерінің қалыптары | | Бір терминнің әрқандай нұсқа топтары |

Шығару рәсімдерінің әрқайсысы өз жұмысы кезінде негізгі getMatches рәсімге жүгінеді, ол Р қалып денесін Т мәтінге қабаттастыруды жүзеге асырады, яғни Т мәтінде Fj барлық үзінділерін анықтау, олардың әрқайсысы қарастырылып отырған Р қалыппен сипатталатын тілдік құралымның нақты қолданылуын білдіреді. Әрбір Fj үзінді (Т мәтін сияқты) қарапайым ti -сөз тұлғалары үзінділерінің белгілі бір тізбегі және оларды бөлетін белгілер болып табылады. Мысалы, [A1] N1 қалыбын мына

*Талдау конструктивті модель негізінде жүргізіледі* сөйлемшесіне қабаттастыру кезінде F1, F2 , F3, F4 төрт үзіндісі табылады: алғашқы үш үзінді бір сөз тұлғасынан (N1 зат есімі) тұрады, ал F4= (t5, t6) үзіндісі сын есім мен зат есімнің тізбегі (A1 N1) - 3.2 суретті қараңыз.



Сурет 3.2. Қалыптардың [ A1 ] N1 қабаттасу нәтижелері

Лексикалық-синтаксистік қалып тек денеден ғана емес, сонымен қатар алу қалыптарын да қамтуы мүмкін болғандықтан, жалпы жағдайда getMatches (Р, Т) рәсімінің жұмысы екі кезеңде жүзеге асырылады:

1) Fj мәтіндік үзінділерін ала отырып, P үлгісінің денесін T мәтініне қабаттастыру;

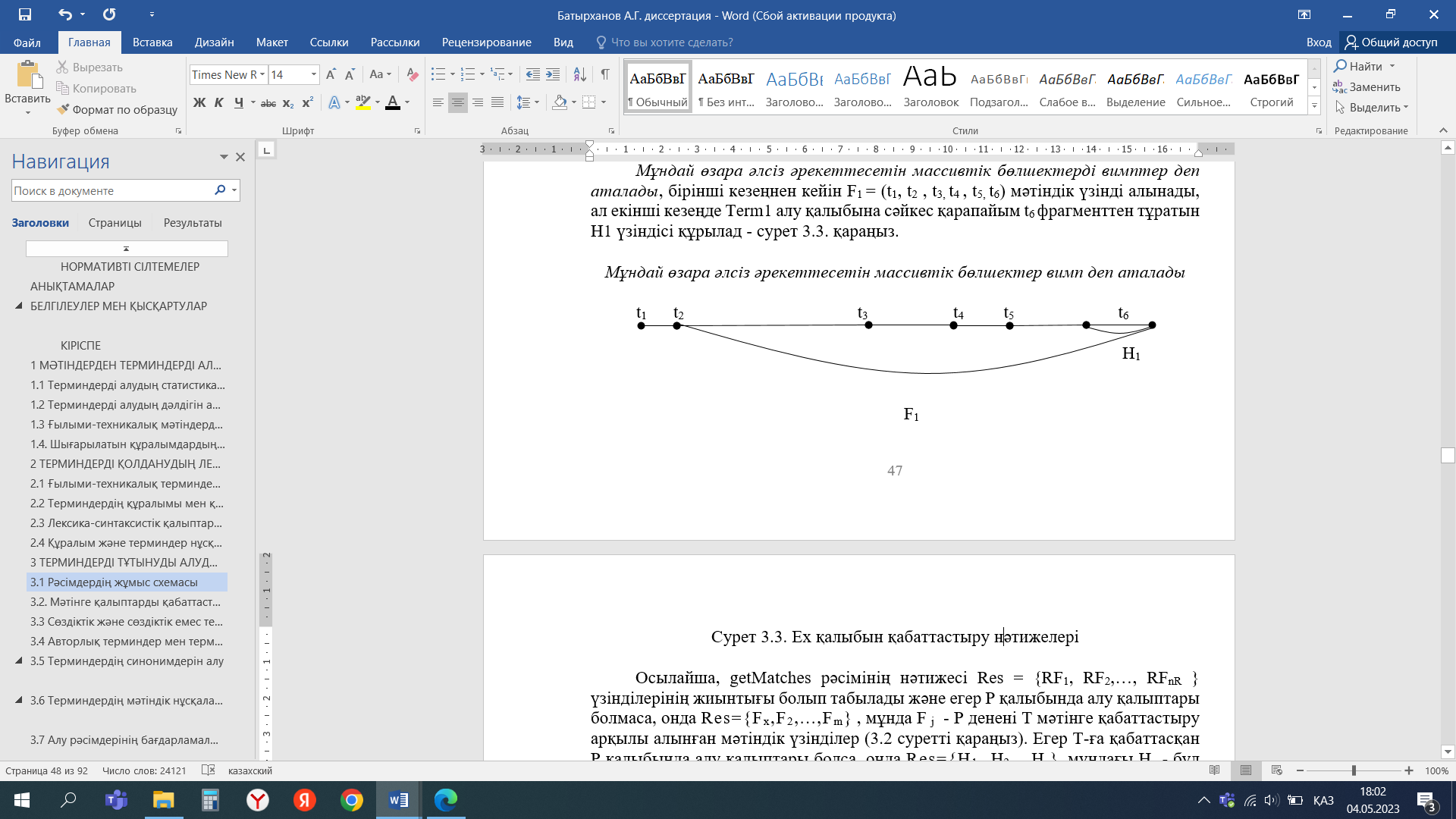
2) Fj мәтіндік үзінділерге Р-дан алу қалыптарына сәйкес Нк үзіндісіне негізделген тізілім құру.

Осылайша, терминнің анықтамалық қалыптарын

Ex = Defin1<c=acc> «атайды»Term1<c=ins># Term1

төмендегі сөйлемшеге қабаттастыру кезінде

*Мұндай өзара әлсіз әрекеттесетін массивтік бөлшектерді вимптер деп аталады*, бірінші кезеңнен кейін F1 = (t1, t2 , t3, t4 , t5, t6) мәтіндік үзінді алынады, ал екінші кезеңде Term1 алу қалыбына сәйкес қарапайым t6 фрагменттен тұратын H1 үзіндісі құрылад - сурет 3.3. қараңыз.



Сурет 3.3. Ex қалыбын қабаттастыру нәтижелері

Осылайша, getMatches рәсімінің нәтижесі Res = {RF1, RF2,…, RFnR } үзінділерінің жиынтығы болып табылады және егер P қалыбында алу қалыптары болмаса, онда Res={Fx,F2,...,Fm}, мұнда Fj - Р денені Т мәтінге қабаттастыру арқылы алынған мәтіндік үзінділер (3.2 суретті қараңыз). Егер Т-ға қабаттасқан Р қалыбында алу қалыптары болса, онда Res={H1, Н2,...,Нр}, мүндағы Нк - бұл алу қалыптарына сәйкес қалыптасқан үзінділер.

Әрі қарай, терминдерді тұтыну рәсімдерінің әрқайсысы (getDictTerms, getNonDictTerms және т.б.) addinTSet көмекші рәсімін тудырады, ол MR={mr1, mr2,mrnM} алынған үміткер-терминдердің жазбаларын құрайтын жиынтық қалыптастырады.

MR жиынтығының әрбір mr1 жазбасының негізгі term1 өрісі алынған терминді қамтиды және getMatches рәсімімен қайтарылған RF1 үзінді бойынша RF1-ге кіретін барлық сөз тұлғаларын бастапқы пішінге келтіру арқылы құрылады (яғни RF1 қалыпқа келтіру арқылы). Қалған жазу өрістерінде term1 үшін қосымша терминологиялық ақпарат жазылады:

* term1 сипаттайтын қалып;
* Т мәтіндегі RF1 үзіндісінің beg шекарасы және end;
* Т –дағы term1 тұтыну жиілігі.

Мұнда beg және end- RF1 басталатын және сәйкесінше аяқталатын қарапайым үзіндінің реттік нөмірі (яғни, үміткер-терминінің Т мәтініне нақты кіруі шекаралармен жазылады). Осылайша, 3.2-суреттегі мысалдағы RF1=F1 мәтіндік үзіндісінің 1 және 1 шекаралары, ал RF4=F4 үзіндісінің 5 және 6 шекаралары бар.

addinTSet рәсімін орындау барысында жаңа RF1 өңдеу кезінде алдымен MR жазбасы бар-жоғы тексеріледі, ол қалыпқа келтірілген RF1 үзіндісімен сәйкес келетін term1 өрісі. Егер мұндай жазба бұрыннан бар болса, оған тек RF1-ге сәйкес келетін beg және end жұбы қосылады. Басқа жағдайда, MR термині үшін жаңа mr1 жазба жасалады, қолдану жиілігі терминдік тұтынуды алудың көлемді рәсіміне байланысты 0 немесе 1-ге тең болады.

Жұмыстың соңында getFrequency көмекші рәсім арқылы алынған алу рәсімдердің әрқайсысына үміткер-терминдерді тұтыну жиілігі есептеледі, санау бір терминнің екіншісіне толық кіргізу жағдайларын ескере отырып жүзеге асырылады. Мысалы, егер мәтінде *мекенжай - логикалық мекенжай* деген бірнеше термин қолданылса, онда *логикалық мекенжай* терминін қолдану жиілігі оның мәтінге кіру саны болып саналады. *Мекенжай* терминін қолдану жиілігі - егер ол *логикалық мекенжай* терминінің бөлігі болмаған жағдайдағы терминнің мәтініне кіру саны.

Осылайша, F1  -ді – term1 терминінің тұтыну жиілігін санау келесі формула бойынша жүзеге асырылады:

F1, = NT1-NP1

Мұнда NTi - term1 терминінің Т мәтінге кіруінің жалпы саны, aл NPl - term1  терминінің басқа терминдердің қандай да бір кіруінің бөлігі болып табылатын кіру саны.

GetFrequency рәсімне кіру кезінде MR жиынтығы келеді, шығу кезінде - MR жиынтығы, онда әрбір term1 үшін Т- дағы оны пайдалану жиілігі көрсетілген. Тек getFrequency рәсімінде мәтінге үміткер-терминдердің кіру шекаралары туралы ақпарат қолданылады.

# 3.2. Мәтінге қалыптарды қабаттастыру ережелері

Барлық алу рәсімдерінің жұмысы getMatches рәсіммен орындалатын Р лексикалық-синтаксистік қалыпты Т өңделген мәтінге қабаттастыруға негізделген. Жалпы жағдайда Р қалыбы мыналарды қамтиды:

* PN – Р қалып атауы (болмауы мүмкін);
* РB - Р қалып денесі, оның е1: РB= (е1,е2,...,еnB) элементтерінің тізбегі болып табылады; жалпы жағдайда әрбір ех үшін (сияқты) x1 морфологиялық сипаттамалар жинағы анықталған;
* РC – бұл е1: РC={с1, с2,..., сnC} қойылатын шарттар (келісу шарттары қалып элементтерінің тиісті морфологиялық сипаттамаларының мәндерінің теңдігін талап етеді, сөздік шарттары қалып элементтерінің сөздікке енуін тексеруді талап етеді);
* Е1,..., ЕnE - алу қалыптары.

Т= (t1, t2,..,tnT) мәтінге Р қалыбының қабаттастыру нәтижесін атайық оның үзіндісі төмендегідей:

а) F қиылыспайтын F1үзінділеріне бөлінеді: F=1=1,nF1, мөлшері қалып элементтерінің санына тең болғанда : n=nB , және F1  әр үзіндісі Р қалыбының е1  элементінің F-ке қабаттастыру нәтижесі болып табылады, және

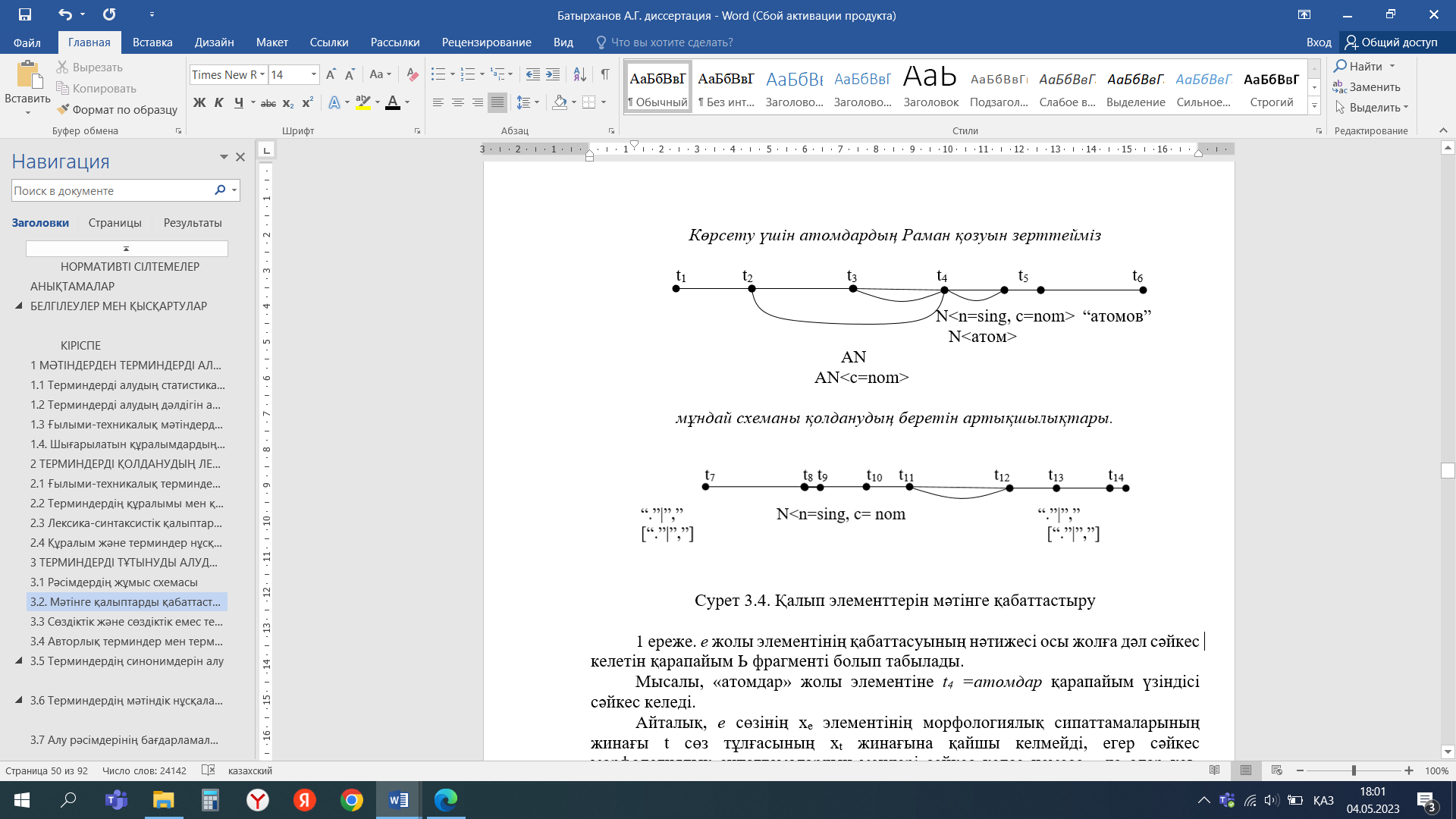
б) F үшін барлық Рс шарттары орындалды, яғни:

* егер сJ – қалып элементтері және е1 және ek үшін сәйкестендіру шарты болса, онда бұл шарт F1 и Fк сәйкес үзінділер үшін орындалады;
* егер сJ, - е1-ге қолданылатын сөздік шарт болса, онда ол сәйкес F1 үшін орындалады.

Осылайша, Т мәтінге Р қалыптың қабаттасуы, Р қалыптың РB денесінің әрбір элементінің дәйекті қабаттасуына және РC шарттарын тексеруге дейін азаяды. Жалпы жағдайда, Т мәтінге Р қалыптың қабаттастыру нәтижелері бірнеше болуы мүмкін - 3.2-суреттегі мысалды қараңыз.

Біз getMatches рәсімі жұмыс істейтін қалып элементтерінің әртүрлі түрлерін қабаттастыру ережелерін сипаттаймыз; біз оларды 3.4-суретте келтірілген мәтіннің мысалында суреттейміз. Жалпы жағдайда, E элементінің нәтижесі қарапайым мәтін үзінділерінің кейбір, мүмкін бос тізбегі болып табылады.

e элементінің қабаттасуы қарапайым мәтін үзінділерінің кейбір, мүмкін бос тізбегі болып табылады.



Сурет 3.4. Қалып элементтерін мәтінге қабаттастыру

1 ереже. *е* жолы элементінің қабаттасуының нәтижесі осы жолға дәл сәйкес келетін қарапайым Ь фрагменті болып табылады.

Мысалы, «атомдар» жолы элементіне t4 =атомдар қарапайым үзіндісі сәйкес келеді.

Айталық, *е* сөзінің хе элементінің морфологиялық сипаттамаларының жинағы t сөз тұлғасының хt жинағына қайшы келмейді, егер сәйкес морфологиялық сипаттамалардың мәндері сәйкес келсе немесе *е*-де олар кез-келген рұқсат етілген мәндерді қабылдай алатындығы анықталған болса. Мысалы, N1<с=ас > сөзі элементі үшін - (с=ас) *табыс септігі* бекітілген, ал сан, тек және жан қатыстылығы кез келген рұқсат етілген мәндерді қабылдай алады. Демек, оның морфологиялық сипаттамалар жинағы сезбек сөз тұлғасының мүмкін болатын мәндері бар атау және табыс септігі бойынша сипаттамалар жинағына қайшы келмейді.

2-ереже. *е* сөзінің элементінің қабаттасуының нәтижесі - сөз тұлғасы болып табылатын 1: қарапайым фрагменті, сонымен бірге:

а) *t* және *e* сөз таптары сәйкес келеді;

б) хе -нің морфологиялық сипаттамаларының жинағы хt жинағына қайшы келмейді.

Мысалдарды қарастырайық. N<n=sing, с=nom> элементті мәтінге қабаттастыру нәтижелері (жекеше атау септігіндегі зат есім) - мысал (3.4-суретті *қараңыз) t3*- *қоздыру және* t11= *қолдану* қарапайым үзінділері болып табылады. Қабаттасу N < атом, n=sing>( *атом* жекеше зат есім) ешқандай нәтиже бермейді, өйткені қарастырылып отырған мәтінде сипатталған шектеулері бар үзінділер жоқ. Сонымен қатар, элементтің қабаттасуының нәтижесі - N<атом> сөзі (*атом* зат есімнің кез келген сөз тұлғасын орнататын) t4 үзіндісі болып табылады.

3-ереже.

а) e=alt1| alt2|… altn түрі қалыптарының балама элементін қабаттастырудың нәтижесі ретінде оның кез келген баламасының аltк қабаттасуының нәтижесі болып табылады.

б) e ={alt1| alt2|… altn} түрінің міндетті элементін қабаттастырудың нәтижесі ретінде кез келген аltк баламасының қабаттасуының нәтижесі болып табылады.

Мысалы, қарастырылып отырған мәтінге қабаттасу нәтижелері « . « | « балама элементтің мысалы t14 үзіндісі (нүкте белгісі), сондай-ақ t8 үзіндісі (үтір белгісі) болып табылады.

4 ереже. Опционалды е=[ор] элементті қабаттастырудың нәтижесі немесе *ор* элементті қабаттастырудың нәтижесі немесе бос үзінді болып табылады.

Мысалы, мәтінге қабаттасудың нәтижесі - [ « :»] опционалды элементінің мысалы - бос үзінді болып табылады (қарастырылып отырған мәтінде қос нүкте белгісі жоқ), бірақ [« . « | «, « ] опционалды элементі қабаттасуының нәтижесі - t8 және t14 үзінділері, сондай-ақ бос үзінді.

5 ереже. е1 е2 ... еn элементтерінің тізбегін қабаттастыру нәтижесі F1F2 ... Fn, где Fi үзінділерінің тізбегі болып табылады, мұндағы Fi өзі еi қабаттасуының нәтижесі.

Мысалы, мәтінге қабаттасудың нәтижесі - «, « Pn1 реттіліктің мысалы - t8 и t9.үзінділерінің тізбегі болып табылады.

6 ереже. Ре деп аталатын қалып данасын қабаттастырудың нәтижесі осы қалыптың денесін қабаттастырудың нәтижесі болып табылады.

Мысалдарды қарастырайық. АN = А N <А=N> (N) қалып анықталсын делік (грамматикалық тұрғыдан бір-біріне сәйкес келетін сын есім мен зат есімнің тізбегі). Осы қалыптың АN данасын қарастырылып отырған мысал мәтініне қабаттастырудың нәтижесі(3.4-суретті қараңыз) F=(t2,t3) үзіндісі болады. Бұл үзінді сонымен қатар АN<с=nom> (А және N атау септігінде сәйкестендірілген) данасын қабаттастырудың нәтижесі болып табылады. Сонымен қатар, AN<c=ins>қалыбының данасын қабаттастыру сәтсіз, өйткені қарастырылып отырған мәтінде көмектес септігінде сәйкестендірілген А және N жоқ.

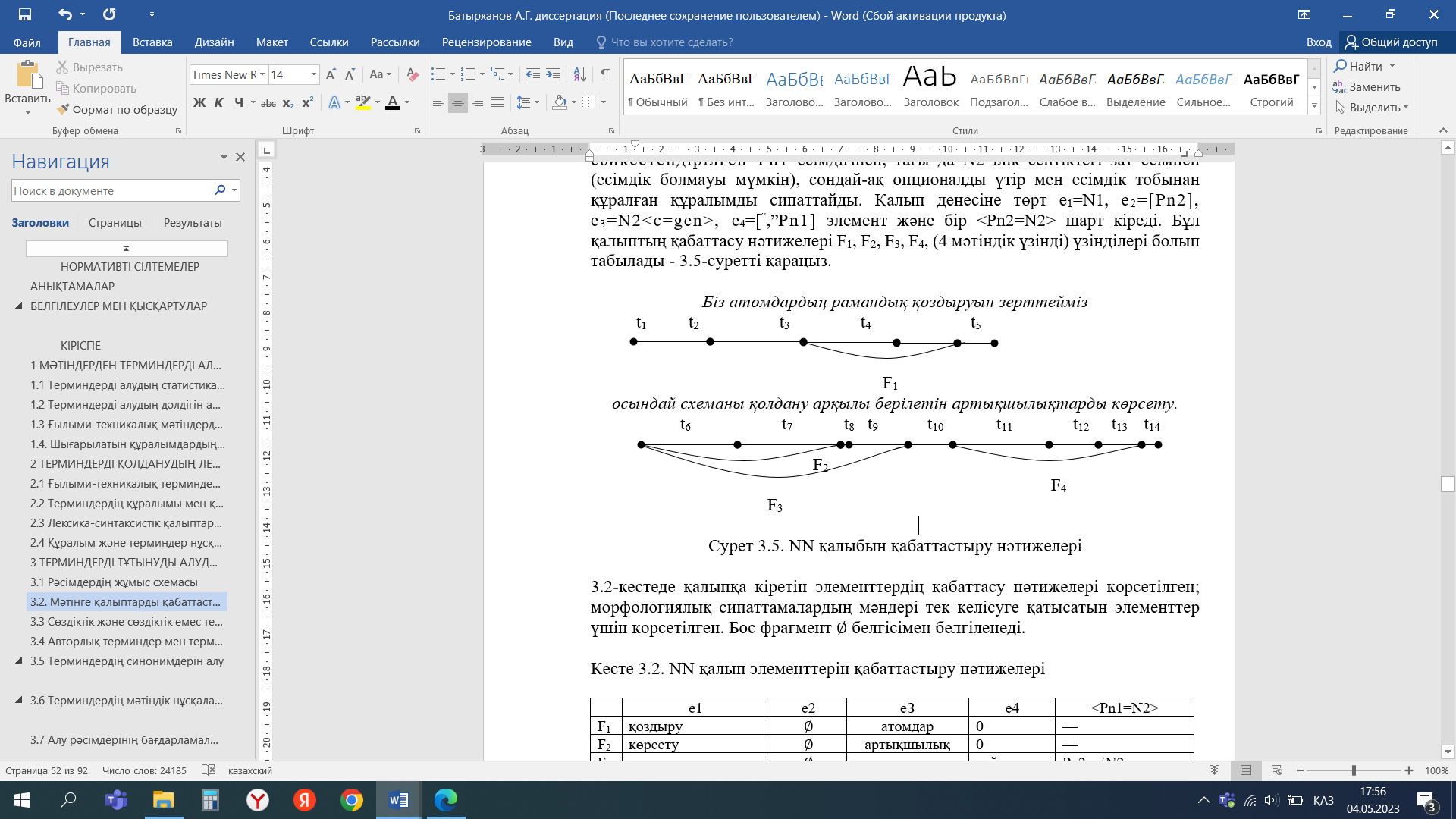
7 ереже. Егер P қалыбында Е1,..., ЕnE алу қалыптары болса, онда F мәтін үзіндісі негізінде - Р қалып денесінің қабаттасуының нәтижесіне сәйкес Еi -ге кіретін элементтер бойынша Нi мәтіндік үзіндісі құрылады.

Мысалы, 3.4-суретте көрсетілген мәтін және VI N1 # N1 қалыптары үшін (оның денесі етістік пен зат есімнің реттілігін орнатады, ал алу қалыбы - зат есім) қабаттасудың нәтижесі F=(t10, t11) – ток өткізу мәтіндік үзінді болады және алу қалыбына сәйкес Н=(t11) - ток мәтіндік үзіндісі құрылады. V1 AN1 # AN1 қалыпты қолданғанда қабаттасудың нәтижесі F=(t1,t2,t3) - раманның қоздыруын зерттейміз үзіндісі болады, алу қалыбына сәйкес H=(t2,t3,) - раманның қоздыруы – үзіндісі пайда болады.

Келесі мысалда қайта сол мәтінге қабаттасуы орын алады - қалыптың мысалы (3.4-сурет)

NN = N1 [Pn1] N2<c=gen> [«, » Pn2] <Рn1=N2>

Бұл қалып N1 зат есімнен және оған жалғасқан грамматикалық сәйкестендірілген Pn1есімдігінен, тағы да N2 ілік септіктегі зат есімнен (есімдік болмауы мүмкін), сондай-ақ опционалды үтір мен есімдік тобынан құралған құралымды сипаттайды. Қалып денесіне төрт e1=N1, е2=[Pn2], е3=N2<с=gen>, е4=[“,”Pn1] элемент және бір <Pn2=N2> шарт кіреді. Бұл қалыптың қабаттасу нәтижелері F1, F2, F3, F4, (4 мәтіндік үзінді) үзінділері болып табылады - 3.5-суретті қараңыз.



Сурет 3.5. NN қалыбын қабаттастыру нәтижелері

3.2-кестеде қалыпқа кіретін элементтердің қабаттасу нәтижелері көрсетілген; морфологиялық сипаттамалардың мәндері тек келісуге қатысатын элементтер үшін көрсетілген. Бос фрагмент ∅ белгісімен белгіленеді.

Кесте 3.2. NN қалып элементтерін қабаттастыру нәтижелері

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | е1 | е2 | еЗ | е4 | <Pn1=N2> |
| F1 | қоздыру |  | атомдар | 0 | — |
| F2 | көрсету |  | артықшылық | 0 | — |
| F3 | көрсету |  | артықшылық  n=plur | қайсысы  n=sign | Pn2.n≠N2.n |
| F4 | қолдану | осындай | схемалар |  | Pn1.c=N2.c=gen  Pn1.g=N2.g=fem  Pn1.n=N2.n=sign |

NN қалыбына N2 зат есімі мен Pn2 есімдігінің морфологиялық сипаттамаларының (сан мен тегінің) теңдігінің қосымша шарттарын қосу жаңа қалып береді:

N1 [Pn1] N2<c=gen> [«,»Pn2]

<Pn1=N2, N2.n=Pn2.n,N2.g=Pn2.g>

мәтінге қабаттасқан кезде-мысал нұсқалар саны төрттен үшке дейін азаяды, өйткені F3 мәтіндік үзінді үшін N2 және Pn2 элементтері үшін n грамматикалық санының теңдік шарты орындалмайды.

Енді терминдерді қолдану рәсімдерінің әрқайсысының жұмыс ерекшеліктерін қарастырамыз.

# 3.3 Сөздіктік және сөздіктік емес терминдерді алу

getDictTerms және getNonDictTerms рәсімдер сәйкесінше сөздіктік және сөздіктік емес терминдердің кіруін алуда ұқсас әрекет етеді: олардың әрқайсысының кірісіне Т талданатын мәтін және Ѕ қажетті қалыптар жинағы келеді. Шығу рәсімінен алынған үміткер-терминдері туралы көптеген жазбалар алады, олардың әрқайсысын Т тұтыну жиілігі туралы ақпарат бар.

Екі рәсімнің жұмысы getMatches(Р, Т) негізгі рәсімін шақыру, S жиынындағы әрбір Р қалып үшін және оның қайтаратын мәндерін MR жиынтыққа қосу болып табылады; рәсімдердің соңында getFrequency көмекші рәсім шақырылады:

getDictTerms/ getNonDictTerms

Input: T, S

Output: MR

begin

MR=0;

for each PiS

begin

F=getMatches (Pi, T);

for eachFi F

addinTSet(Fi, MR, 0) ;

end

getFrequency(MR) ;

end

Мысалда getDictTerms рәсімінің жұмысын қарастырамыз. 3.6-суретте көрсетілген мәтін кіріске түссін, ал кіріс жинағы ретінде келесі сөздіктік қалыптар алынады:

Р1: жедел = А 1 < жедел> {Ра 1<есте сақтау>

N1 < құрылғы><A1=A1=N1> / N1<жад><A1=N 1>}

Р2:: жад = N1 < жад> [N2 < пайдаланушы, бірге=gen> |

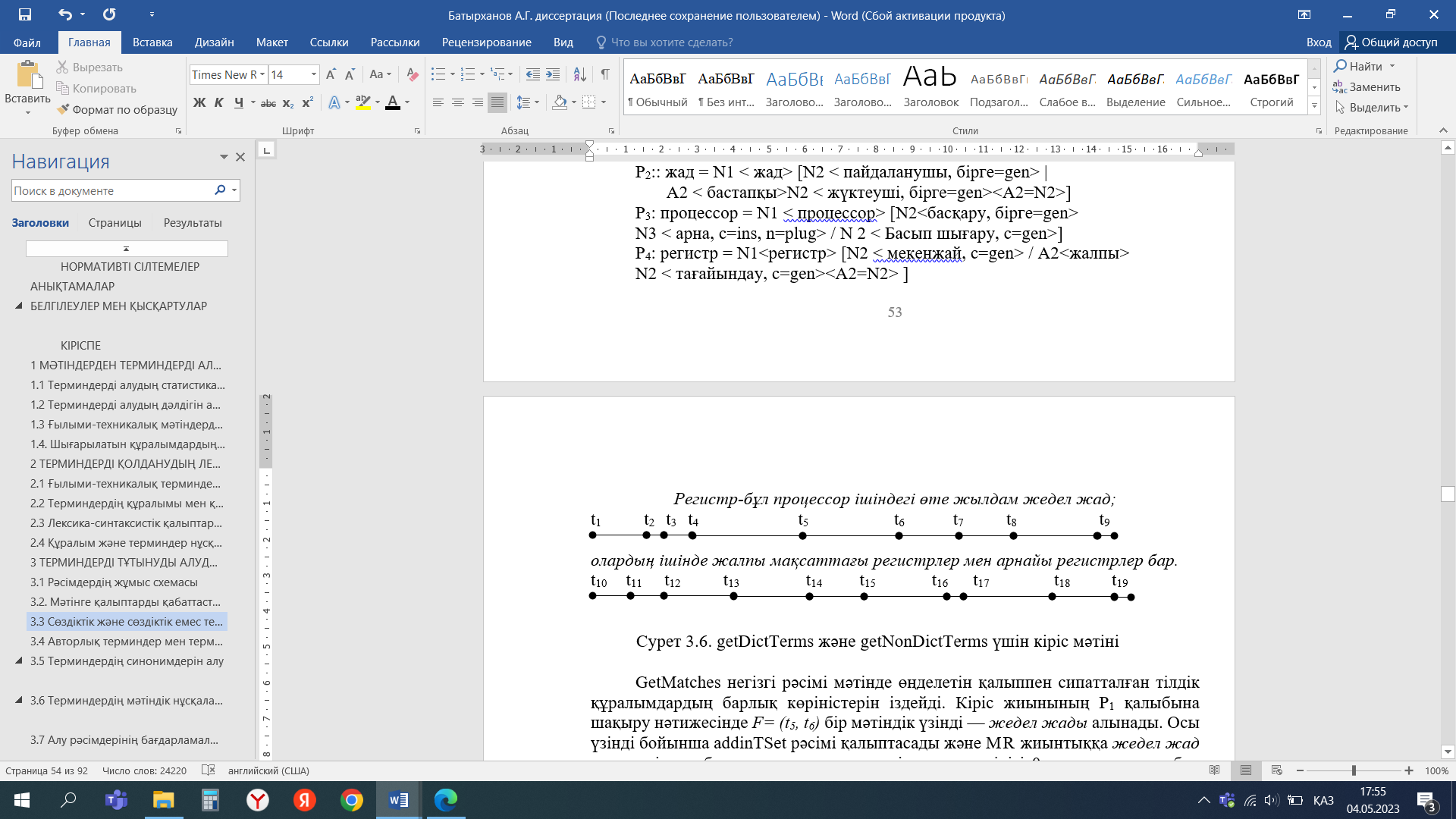
A2 < бастапқы>N2 < жүктеуші, бірге=gen><A2=N2>]

Р3: процессор = N1 < процессор> [N2<басқару, бірге=gen>

N3 < арна, с=ins, n=plug> / N 2 < Басып шығару, с=gen>]

Р4: регистр = N1<регистр> [N2 < мекенжай, с=gen> / А2<жалпы>

N2 < тағайындау, с=gen><A2=N2> ]



Сурет 3.6. getDictTerms және getNonDictTerms үшін кіріс мәтіні

GetMatches негізгі рәсімі мәтінде өңделетін қалыппен сипатталған тілдік құралымдардың барлық көріністерін іздейді. Кіріс жиынының Р1 қалыбына шақыру нәтижесінде F= (t5, t6) бір мәтіндік үзінді — *жедел жады* алынады. Осы үзінді бойынша addinTSet рәсімі қалыптасады және МR жиынтыққа *жедел жад* терминнің жазбасы қосылады, терминді тұтыну жиілігі 0-ге орнатылады, бұл addinTSet рәсімінің үшінші аргументін белгілейді.

addinTSet рәсімінің үшінші дәлелі мынаны көрсетеді:

* термин туралы жаңа жазбаның МR жиынтығына енгізген кезде терминді тұтыну жиілігі қандай мәнге орнатылуы керек;
* терминді тұтыну жиілігі оның МR жиынтығына кіруінің шекараларына қосылған кезде қандай мәнге артуы керек. Үшінші аргумент әр түрлі алу рәсімдерінде әр түрлі мағынаға ие болуы мүмкін екенін ескереміз. Сипатталған рәсімдерден addinTSet шақырған кезде аргументтің нақты мәні 0-ге тең болады, өйткені терминдерді бір-біріне енгізу мүмкіндігіне байланысты терминдердің жиіліктерінің нақты мәндерін тек getFrequency рәсімі арқылы табуға болады, содан кейін ол осыны жасайды.

Қабаттасу нәтижелері және басқа қалыптар да осылай өңделеді. Р4 қалыбын қабаттастыру кезінде төрт қабаттасу нәтижесі табылатынын ескереміз: *регистр* термині үшін үш мәтіндік үзінді және *жалпы мақсаттағы регистр* термині үшін біреуі.

Қалыптардың кіріс жинағын өңдеу нәтижесінде алынған МR жиынтығы 3.3-кестеде келтірілген, соңғы бағанда getFrequency рәсімін орындағаннан кейін алынған үміткер-терминдердің жиіліктерінің қорытынды мәндері көрсетілген. Басқа терминдердің кіру бөлігі болатын терминдердің кіру шекаралары қалың әріппен белгіленген.

Кесте 3.3. getDictTerms рәсімінің жұмыс нәтижесі

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № | Үміткер-термин | Терминді сипаттайтын қалып | Термин шекарасы | Термин жиілігі |
| 1. | Жедел жад | А1<жедел>N1<жад><А1=N1> | 5, б | 1 |
| 2. | жад | N1<жад> | 6,6 | 0 |
| 3. | процессор | N1<процессор> | 8,8 | 1 |
| 4. | регистр | N1<регистр> | 1,1  13,13  18,18 | 2 |
| 5. | жалпы мақсаттағы регистр | N1<регистр>A2<жалпы>  N2<мақсаттағы, c=gen>  <A2=N2> | 13,15 | 1 |

Әрбір үміткер-термині үшін getFrequency рәсімді орындау барысында оның кіруі басқа терминдердің мәтініне кірудің бір бөлігі болып табылатындығы тексерілді. *Жедел жад* терминінің кіруі басқа терминдердің кіруінің бөлігі емес, сондықтан оны қолдану жиілігі 1-ге тең; дәл осылай *процессор* және *жалпы мақсаттағы регистр* терминдері үшін жиілік есептелді. Сонымен қатар, *жад* терминінің кіруі *жедел жад* терминінің кіруіне енгізіледі (F=(t6,t6) мәтіндік үзінді F=(t5,t6) үзіндінің бөлігі болып табылады), сондықтан *жад* терминін қолдану жиілігі жоғарылаған жоқ және 0-ге тең болып қалды. *Регистр* термин мәтініндегі үш кірістің бірі үшін оның басқа терминге енуі де бар, сондықтан бұл терминді қолдану жиілігі 2-ге тең.

Енді сол мәтіннің мысалында (3.6-сурет) getNonDictTerms рәсімінің жұмысын қарастырамыз. Біріншіден, getMatches рәсімінің жұмысында мәтінде сөздердің және сөз тіркестерінің барлық кірістері ізделеді, олар қалыппен қарастырылған синтаксистік үлгілерге ЅР деген атаулы ескерілетін қалыптарымен сәйкес келеді (қалып ҒТ мәтіндеріне тән синтаксистік терминдердің үлгілерін ескереді). Қарастырылып отырған мәтін үшін қабаттасудың он бір нәтижесі алынады, олар бойынша addinTSet рәсімімен алынған үміткер-терминдері туралы тоғыз жазба жасалады - 3.4 - кестені қараңыз. Кестенің соңғы бағанында getFrequency рәсімімен орнатылған жиіліктің қорытынды мәндері көрсетілген. Терминдердің кірістірілген пайда болуының шекаралары қалың әріппен белгіленген.

Үміткер-терминдерді қолдану жиілігін санау кезінде getFrequency рәсімі кейбір терминдердің басқаларына кірігуі табылды, мысалы, *регистр* термині *арнайы регистр* және *жалпы мақсаттағы регистр* терминдерінің бөлігі болып табылады. Жиілікті есептеу кезінде кірігулер есепке алынбағандықтан, үміткер-терминдер үшін *жад, мақсат, жедел жад* және *жалпы мақсат* қолдану жиілігі 0-ге тең болып қалды.

Кесте 3.4. getNonDictTerms рәсімінің жұмыс нәтижесі

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № | Үміткер-термин | Терминді сипаттайтын қалып | Термин шекарасы | Термин жиілігі |
|  |
| 1. | *регистр* | N1<регистр> | 1,1  13,13  18,18 | 1 |
| 2. | *жад* | N1<жад> | 6,6 | 0 |
| 3. | *процессор* | N1<процессор> | 8,8 | 1 |
| 4. | *мақсат* | N1<мақсат> | 15,15 | 0 |
| 5. | *Жедел жад* | A1<жедел>N1<жад><А1=N1> | 5,6 | 0 |
| 6. | *Жалпы мақсат* | А1<жалпы>  N1<мақсат><А1=N1> | 14,15 | 0 |
| 7. | *арнайы регистр* | А1<арнайы>N1<регистр><А1=N1> | 17,18 | 1 |
| 8. | *Шапшаң жедел жад* | А1<шапшаң> А2<жедел>N1<жад><А1=А2=N1> | 4,6 | 1 |
| 9. | *жалпы мақсаттағы регистр* | N1<регистр>А2<жалпы>N2<мақсат, бірге-gen><А2=N2> | 13,15 | 1 |

getDictTerms және getNonDictTerms рәсімдері арқылы алынған үміткер-термин жиынтықтарында қиылысу бар екенін байқаймыз - бұл *регистр, процессор* және *жалпы мақсаттағы регистр* терминдері.

# 3.4 Авторлық терминдер мен терминдерді қосылыстардан алу

Авторлық терминдерді алудың getAuthTerms рәсімі және терминдерді қосылыстардан алудың getTermsfromCombs рәсімі ұқсас жұмыс істейді: әр рәсімнің кірісіне талданатын Т мәтін және S қалыптарының қажетті жинағы келеді. Шығу кезінде Т мәтінінде олардың әрқайсысын қолдану жиілігі туралы ақпараты бар үміткер-терминдер туралы жазбалар МR жиынтығы алынады. Терминдерді алу келесідей жүреді: алдымен кіріс қалыптарымен танылған құралымдардан терминдер алынады, содан кейін осы терминдердің мәтінге барлық кірістері ізделеді.

getAuthTerms / getTermsfromCombs

Input: T, S

Output: MR

begin

MR=0;

for each Pi S //қадам 1

begin

F=getMatches(Pi, T) ;

for each Fi F

addinTSet (Fi, MR,1);

end

Sext=makePSet (MR) ; //қадам 2

for each PiSext //қадам 3

begin

F=getMatches (Pi, T);

for each Fi F

addinTSet (Fi,MR, 0) ;

end

getFrequency(MR); //қадам 4

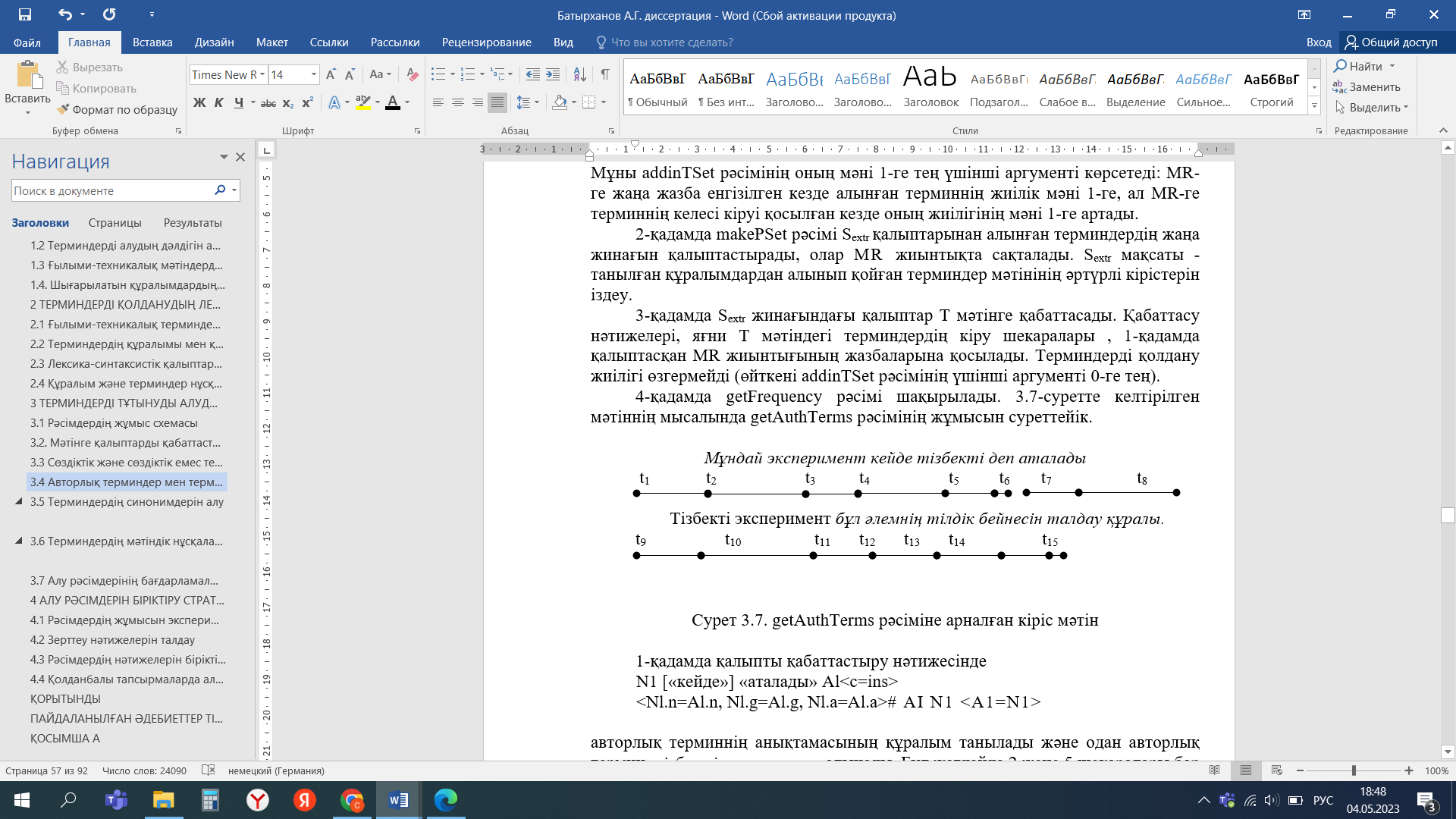
end

Осы рәсімдердің 1-қадамында getMatches негізгі рәсімін қолдана отырып (2 кезеңде жұмыс істейді) кіріс қалыптары бойынша терминдер алынады, алынған терминдер үшін addinTSet рәсімі МR жиынтығын құрайды. Қарастырылып отырған рәсімдермен танылған құралымдарда терминдердің үзілген қолданылуы мүмкін болғандықтан - 3.7 және 3.8 суреттерді қараңыз, 1-қадамда табылған әрбір терминнің кіруі бірден MR жиынтығында ескеріледі. Мұны addinTSet рәсімінің оның мәні 1-ге тең үшінші аргументі көрсетеді: MR-ге жаңа жазба енгізілген кезде алынған терминнің жиілік мәні 1-ге, ал MR-ге терминнің келесі кіруі қосылған кезде оның жиілігінің мәні 1-ге артады.

2-қадамда makePSet рәсімі Sextr қалыптарынан алынған терминдердің жаңа жинағын қалыптастырады, олар MRжиынтықта сақталады. Sextr мақсаты - танылған құралымдардан алынып қойған терминдер мәтінінің әртүрлі кірістерін іздеу.

3-қадамда Sextr жинағындағы қалыптар Т мәтінге қабаттасады. Қабаттасу нәтижелері, яғни Т мәтіндегі терминдердің кіру шекаралары , 1-қадамда қалыптасқан MR жиынтығының жазбаларына қосылады. Терминдерді қолдану жиілігі өзгермейді (өйткені addinTSet рәсімінің үшінші аргументі 0-ге тең).

4-қадамда getFrequency рәсімі шақырылады. 3.7-суретте келтірілген мәтіннің мысалында getAuthTerms рәсімінің жұмысын суреттейік.



Сурет 3.7. getAuthTerms рәсіміне арналған кіріс мәтін

1-қадамда қалыпты қабаттастыру нәтижесінде

N1 [«кейде»] «аталады» Al<c=ins>

<Nl.n=Al.n, Nl.g=Al.g, Nl.a=Al.a># AI N1 <A1=N1>

авторлық терминнің анықтамасының құралым танылады және одан авторлық термин *тізбекті эксперимент* алынады. Бұл жағдайда 2 және 5 шекаралары бар анықталған терминнің кіруі (оның үзілгенін ескеріңіз) MR жиынтығына қосылады, терминді қолдану жиілігі 1-ге тең болады.

2-қадамда алынған терминді сипаттайтын қалып кіретін Sextr жинағы құрастырылған:

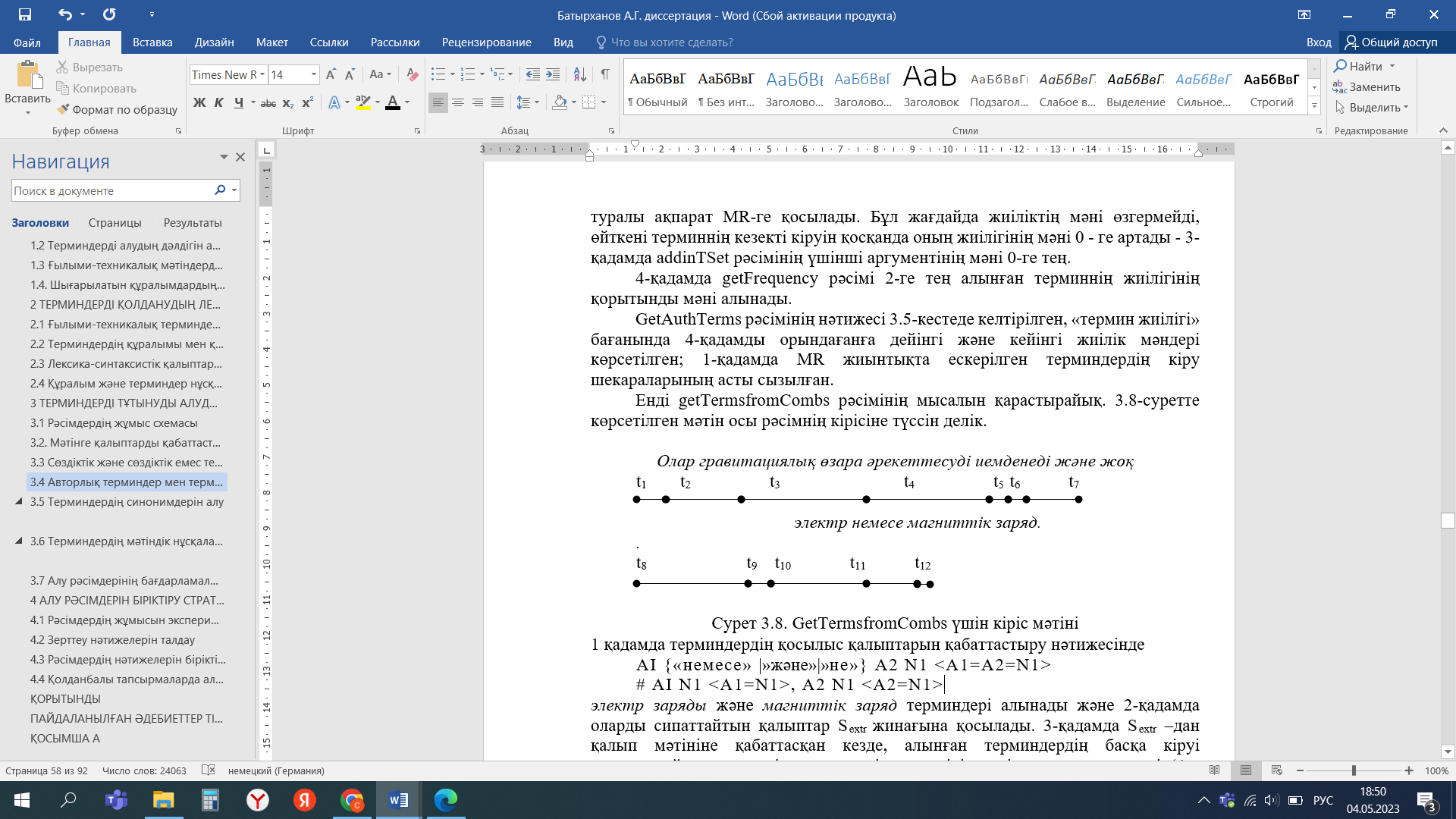
A1 < тізбек>N1 < эксперимент><A1=N1>

3-қадамда Sextr–ден қалып мәтініне қабаттасқан кезде, 7 және 8 шекаралары бар, *тізбекті эксперимент* терминнің екінші кіруі табылады - бұл туралы ақпарат MR-ге қосылады. Бұл жағдайда жиіліктің мәні өзгермейді, өйткені терминнің кезекті кіруін қосқанда оның жиілігінің мәні 0 - ге артады - 3-қадамда addinTSet рәсімінің үшінші аргументінің мәні 0-ге тең.

4-қадамда getFrequency рәсімі 2-ге тең алынған терминнің жиілігінің қорытынды мәні алынады.

GetAuthTerms рәсімінің нәтижесі 3.5-кестеде келтірілген, «термин жиілігі» бағанында 4-қадамды орындағанға дейінгі және кейінгі жиілік мәндері көрсетілген; 1-қадамда MR жиынтықта ескерілген терминдердің кіру шекараларының асты сызылған.

Енді getTermsfromCombs рәсімінің мысалын қарастырайық. 3.8-суретте көрсетілген мәтін осы рәсімнің кірісіне түссін делік.



Сурет 3.8. GetTermsfromCombs үшін кіріс мәтіні

1 қадамда терминдердің қосылыс қалыптарын қабаттастыру нәтижесінде

AI {«немесе» |»және»|»не»} А2 N1 <A1=A2=N1>

# AI N1 <A1=N1>, А2 N1 <A2=N1>

*электр заряды* және *магниттік заряд* терминдері алынады және 2-қадамда оларды сипаттайтын қалыптар Sextrжинағына қосылады. 3-қадамда Sextr –дан қалып мәтініне қабаттасқан кезде, алынған терминдердің басқа кіруі анықталмайды және екі термин үшін де жиіліктердің қорытынды мәні (4 - қадамнан кейін) 1-ге тең болады -3.5-кестені қараңыз.

Кесте3.5. getAuthTerms және getTermsfromCombs нәтижелері

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Үміткер-термин | Терминді сипаттайтын қалып | Термин шекарасы | Термин жиілігі | |
| getAuthTerms рәсімімен жұмыс нәтижесі | | | | |
| *Тізбекті эксперимент* | А1<тібекті>  N1<эксперимент><А1=N1> | 2,5  7,8 | 11 | 22 |
| getTermsfromCombs рәсімімен жұмыс нәтижесі | | | | |
| *электр заряды* | А1<электр>  N1<заряд><А1=N1> | 8,11 | 11 | 11 |
| *магниттік заряд* | А1<магниттік>N1<заряд><А1=N1> | 10,11 | 11 | 11 |

1-қадамда анықталған 10 және 11 шекаралары бар магниттік заряд терминінің пайда болуы 8 және 11 шекаралары бар *электр заряды* терминінің кіруіне енгізілгенін ескереміз (3.8-суретті қараңыз). 1-қадамда анықталған терминнің кіруі MR жиынтығында бірден ескерілетіндіктен-олардың әрқайсысы терминнің жиілігін 1-ге арттырады және getFrequency рәсімі жұмыс істеген кезде мұндай пайда болмайды, нәтижесінде бұл терминдердің жиілігі дұрыс мағынаға ие болады.

# 3.5 Терминдердің синонимдерін алу

GetSynTerms енгізу рәсімінде синонимдерді енгізу құралымдарынан терминдерді алу үшін Т талданатын мәтін және Б қалыптар жиынтығы келеді; шығуда төмендегідей болады:

* алынған терминдер туралы MR жазбалары жиынтығы және олардың синонимдері, қолдану жиілігі туралы ақпарат;
* < термин-синоним> жұптарынан алынған жинақ;
* шығару келесідей жүреді: алдымен S қалып жинақтары бойынша танылған синонимдерді енгізу құралымдарынан терминдер мен олардың синонимдері алынады, содан кейін олардың мәтінге барлық кірістері ізделеді.

getSynTerms

Input: T, S

Output: MR, С

begin

MR= 0;

for each PiS //қадам 1

begin

F=getMatches (Pi, T) ;

makeSynonyms(F,C);

for each FjF

addinTSet (Fi,MR,1) ;

end

Sext=makePSet (MR) ; //қадам 2

for eachPi Sext //қадам З

begin

F=getMatches(Pi, T) ;

for eachFjF

addinTSet (Fi, MR,0);

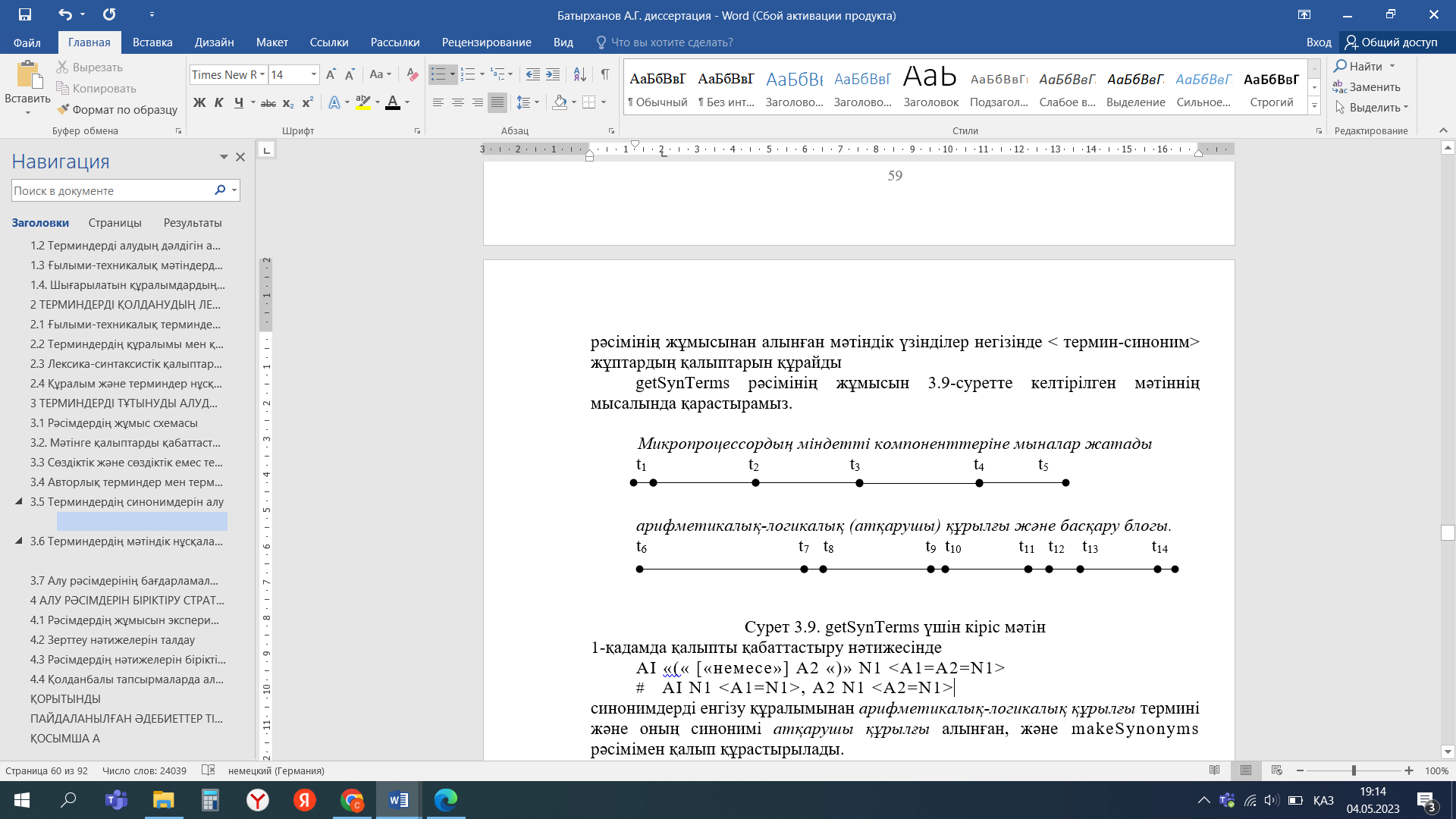
end

getFrequency (MR) ; //қадам 4

end

getSynTerms рәсімінің алдыңғы бөлімде сипатталған алу рәсімінен негізгі айырмашылығы - makeSynonyms рәсімін 1-қадамында шақыру, ол getMatches рәсімінің жұмысынан алынған мәтіндік үзінділер негізінде < термин-синоним> жұптардың қалыптарын құрайды

getSynTerms рәсімінің жұмысын 3.9-суретте келтірілген мәтіннің мысалында қарастырамыз.



Сурет 3.9. getSynTerms үшін кіріс мәтін

1-қадамда қалыпты қабаттастыру нәтижесінде

AI «(« [«немесе»] А2 «)» N1 <A1=A2=N1>

* AI N1 <A1=N1>, А2 N1 <A2=N1>

синонимдерді енгізу құралымынан *арифметикалық-логикалық құрылғы* термині және оның синонимі *атқарушы құрылғы* алынған, және makeSynonyms рәсімімен қалып құрастырылады.

A1<арифметикалық-логикалық>N1 < құрылғы><A1=N1>

# A2 < атқарушы > N2 < құрылғы><A2=N2> *арифметикалық-логикалық құрылғы* және *атқарушы құрылғы* терминдері берілген мәтінде синоним болып табылады.

Мәтіннің синонимдерін сипаттайтын ұқсас қалыптар терминдердің дублеттерін сипаттайтын қалыптармен қатар мәтіндік нұсқаларды шығару үшін кейінірек getVarsforTerms рәсімінде қолданылатындығын ескертеміз,.

2-қадамда алынған терминді және оның синонимін сипаттайтын қалыптар енгізтін Sextr жинағы құрастырылады.

3-қадамда Sextr -дағы мәтін қалыптарын қабаттастыруда терминдердің басқа кірісі анықталмайды және олардың жиіліктерінің қорытынды мәні 1-ге тең болып қалады.

GetSynTerms рәсімінің жұмысынан алынған MR жиынтығы 3.6-кестеде келтірілген. «Термин жиілігі» бағанында 4-қадамды орындағанға дейінгі және кейінгі жиілік мәндері көрсетілген.

Кесте 3.6. getSynTerms рәсімімен анықталған терминдер

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Үміткер-термин | Терминді сипаттайтыын қалып | Термин шекарасы | Термин жиілігі | |
| *Арифметикалық-*  *логикалық құрылғы* | *А1< арифметикалық-логикалық>*  N1<құрылғы><A1=N1> | 6, 10 | 1 | 1 |
| *Атқарушы құрылғы* | А1<атқарушы>N1<құрылғы><A1=N1> | 8, 10 | 1 | 1 |

# 3.6 Терминдердің мәтіндік нұсқаларын тану

GetVarsforTerms рәсімінің кіруіне терминдердің мәтіндік нұсқаларын алу орнатылады:

1) мәтіндік нұсқаларды табуды қажет ететін Mterm терминдері жиынтығы;

2) сөздер мен сөз тіркестерінің Mcand жиынтығы, олардың арасында мәтіндік нұсқаларды іздеу жүзеге асырылады;

3) терминдердің мәтіндік нұсқаларын құру ережелерінің Svar қалыптарының жинағы.

Бір терминнің нұсқалар болып саналатын G1,G2, ..., GN сөздер мен сөз тіркестері топтары рәсімінің шығуы.

getVarsforTerms

Input:Mterm,Mcand, S var

Output: G1, G2, ...,Gn

for each termiMterm

for each termj Mcand

begin

ifisVar (termi, termj, Svar)

thenaddinGroup (Gterm, termj) ;

end

getVarsforTerms рәсімінің жұмысы isVar көмекші рәсімді шақырудан тұрады, ол берілген termx және term - екі сөз/сөз тіркесі үшін, termi  - termj -дің нұсқасы бола алатынын тексереді. Тексеру кезінде негізгі getMatches рәсіміне жүгіну жүзеге асады; мәтіндік нұсқаларды Svar жинағының қалыптарымен беріледі, құру ережелері isVar теру үлгілерімен беріледі - isVar рәсімінің шығуында - мәні true немесе false.

Егер isVar рәсімінің нәтижесі true мәні болса, онда addinGroup көмекші рәсімі шақырылады, ол Gtermi тобына term1 терминіне сәйкес келетін оның termj нұсқасын қосады.

Алдымен getVarsforTerms рәсімінің жұмысын 1-мысалда қарастырамыз - *қолданбалы бағдарламалық жасақтама* және *қоқыс жинау* терминдеріне арналған дублеттерді тану. Рәсімнің кіріс мәліметтері 3.7-кестеде келтірілген.

Кесте 3.7. 1-мысал үшін getVarsforTerms кіріс мәліметтері

|  |  |
| --- | --- |
| Атауы | *Мамұны* |
| Mterm | қолданбалы бағдарламалық жасақтама, қоқыс жинау |
| Mcand | бағдарламалық жасақтама , қолданбалы БЖ , жад тазалау |
| Svar – жұп қалыптары<термин-дублет> | *N1<жиын>N2<қоқыс, с=gen>*  *# N3<тазалау>N4<жад,с=gen>* |

Терминдердің дублеттерін іздеу кезінде isVar рәсімінің жұмысы тек *қоқыс жинау - жадты тазарту* жұптары үшін сәтті болады, өйткені ол үшін Svar жиынтығында тиісті қалып бар; isVar жиынтығында тиісті шаблондардың болмауына байланысты барлық басқа қарастырылған терминдер жұбы үшін (*қолданбалы бағдарламалық жасақтама — қолданбалы БЖ, қолданбалы бағдарламалық жасақтама - жадты тазарту* және т. б.) isVar рәсімі false мәнін береді. Нәтижесінде, getVarsforTerms рәсімінен G= {*қоқыс жинау, жадты тазарту*}нұсқалардың бір тобы құрылады.

Енді 2-мысалда getVarsforTerms рәсімінің жұмысын қарастырмыз - осы терминдердің лексикалық-синтаксистік нұсқаларын тану. Кіріс мәліметтері 3.8-кестеде келтірілген; кестенің соңғы жолында синонимдер сөздігінің үзіндісі көрсетілген, оны getMatches рәсімі Syn(A3, A1) түріндегі қалыптарда көрсетілген сөздік шарттарын тексеру үшін қолданады.

Svar жинағында A1 A2 N1 (*қолданбалы бағдарламалық жасақтама*) түрінің терминдері үшін нұсқалардың пайда болуын сипаттайтын қалып бар: нұсқалар - A2 N1 түрінің тіркестері (*бағдарламалық жасақтама*) және қалыппен сипатталған тіркестер

A1 N2 <Syn (A2 N1, N2), A1=N2>

Бұл қалып бірінші сөз бастапқы - A1 (*қолданбалы*) терминнің бірінші сөзімен сәйкес келетін нұсқаны белгілейді, ал N2 болса A2 N1 (*бағдарламалық жасақтама*) сөз тіркесінің синонимі болып табылады: бұны Syn(A2 N1, N2) сөздік шарты бекітеді. Қарастырылып отырған мысал үшін көрсетілген сөздік шарт орындалады - 3.8-кестенің соңғы жолындағы үлгіні қараңыз.

Кесте 3.8. 2-мысал үшін getVarsforTerms кіріс мәліметтері

|  |  |
| --- | --- |
| Атауы | Мазмұны |
| Mterm | қолданбалы бағдарламалық жасақтама, қоқыс жинау |
| Mcand | бағдарламалық жасақтама , қолданбалы БЖ , жад тазалау |
| Svar- лексикалық-  синтаксистік нұсқалардың жасалу ережелерінің қалыптары | А1 А2 N1 <А1=А2=N1> # |
| N1, А2 N1 <А2=N1>, А1 N1 <А1=N1>, |
| АЗ А2 N1 <Syn (АЗ, А1) , А2=АЗ=N1>, |
| А1 АЗ N1 <Syn(АЗ,А2),А1=АЗ=N1>, |
| А1 А2 N2 <Syn (N1,N2),А1=А2=N2>, |
| А1 N2 <Syn (А2 N1, N2 ) , А1=N2> |
| Syn | А1<бағдарламалық>N1<жасақтама><А1=N1> |
|  | # N2<ПО> |

Осылайша, лексикалық-синтаксистік нұсқаларды тану кезінде isVar рәсімі жұптар үшін true мәнін береді *қолданбалы бағдарламалық жасақтама* - *бағдарламалық жасақтама* және *қолданбалы бағдарламалық жасақтама*-*қолданбалы БЖ*. Нәтижесінде getVarsforTerms рәсімі мәтіндік нұсқалардың келесі тобын алады:

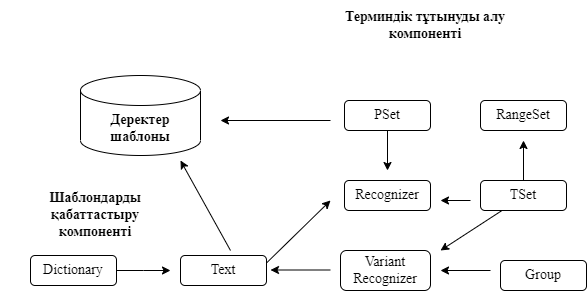
G= {*қолданбалы бағдарламалық жасақтама, қолданбалы бағдарламалық жасақтама, бағдарламалық қамтамасыз ету*}

Көріп отырғаныңыздай, getVarsforTerms рәсімінің схемасы бұрын сипатталған алу рәсімдерінен ерекшеленеді: оған алдын ала құрастырылған Mterm және Mcand. жиынтықтары кіреді. getVarsforTerms рәсімі Т мәтінді кіріске ала алады, бірақ бұл жағдайда оның жұмысының дәлдігі өте төмен болады, өйткені нұсқалар мәтіннің барлық сөздері үшін олардың терминологиялығы мен бір-біріне кірігуін ескермей анықталады.

# 3.7 Алу рәсімдерінің бағдарламалық қамтамасыз ету

Рәсімдерді жүзеге асыру Linux ОЖ платформасында STL және LSPL кітапханаларын қолдана отырып, С++ [тілінде жүргізілді. Жасалған бағдарламалық жасақтаманың архитектурасы 3.10-суретте көрсетілген, оның негізгі компоненттері:

* терминдерді және олардың мәтіндегі әртүрлі кіруін тануды, сондай-ақ терминдер мен (Recognizer кластары, VariantRecognizer, PSet, TSet, Group және RangeSet) нұсқаларының топтары туралы жазбалар жиынтықтарын өңдеуді қамтамасыз ететін терминдерді тұтынуды алу компоненті;
* мәтінді талдау және оған қалыптарды қабаттастыру бойынша негізгі функцияларды орындайтын қалыптарды қабаттастыру компоненті (Text жәнеDictionaryкластары);
* терминдерді, сондай-ақ мәтіндегі құралымдар мен оларды қолдану нұсқаларын сипаттайтын қалыптар базасы.



Сурет 3.10. Бағдарламалық құрал архитектурасы

3.10-суретте әзірленген бағдарламалық сыныптардың байланыстары көрсетілген, 3.9-кестеде олардың негізгі әдістерінің тізімі келтірілген.

Кесте 3.9. Бағдарламалық кластар және олардың әдістері

|  |  |  |
| --- | --- | --- |
| Класс атауы | Класс әдістері | Сипаттама |
| Text — өңделетін мәтіндегі тілдік құрылымдарды тану | | |
|  | getMatches | Қалыпты мәтінге қабаттастыру |
| getNormalizedText | Мәтіндік үзіндіні қалыпқа келтіру |
| getPattern | Мәтіндік үзінді бойынша қалыпты қалыптастыру |
| Dictionary – машиналық сөздіктерді ұсыну (дерексіз класс) | | |
|  | accepts | Сөздікке кіруді тексеру |
| makeSynonyms | сөздікке термин-синоним жұптарын қосу |
| Recognizer – терминдерді тұтынуды алу (дерексіз класс) | | |
|  | RecognizeAndSearch | Терминдерді қолдану (дерексіз әдіс) |
| VariantRecognizer – мәтіндік нұсқаларды тану | | |
|  | isVar | сөз/сөз тіркестерінің жұбы нұсқалар екенін тексеру |
| getVarsforTerms | Нұсқалар тобын қалыптастыру |
| TSet – үміткер-терминдер туралы жазбалар жиынтығымен жұмыс | | |
|  | addinTSet | Жиынтыққа қосу |
| searchinTSet | Жиынтықта іздеу |
| PSet – қалыптар жиынтығымен жұмыс | | |
|  | makePSet | Қалыптар жиынтығын қалыптастыру |
| Group – үміткер-нұсқалар топтарымен жұмыс | | |
|  | addinGroup | Нұсқалар тобына қосу |
| RangeSet- тұтыну жиілігін есептеу | | |
|  | getFrequency | Тұтыну жиілігін есептеу |

Сипатталған алу рәсімдерін жүзеге асыру үшін RecognizeAndSearch негізгі әдісімен Recognizer дерексіз класынан алынған бағдарламалық кластар жасалған:

* DictionaryRecognizer класы - сөздіктік терминдердің кіруін алу (getDictTerms рәсімін жүзеге асыру);
* NonDictionaryRecognizer класы - сөздіктік емес терминдердің кіруін алу ( getNonDictTerms рәсімін жүзеге асыру);
* AuthorRecognizer класы - авторлық терминдерді алу (getAuthTerms рәсімін жүзеге асыру);
* SynonymRecognizer класы - синонимдерді алу (getSynTerms рәсімін жүзеге асыру);
* CombinationRecognizer класы - қосылыстардан терминдерді алу (getTermsfromCombs рәсімін жүзеге асыру).

Мәтіндік нұсқаларды тану рәсімін жүзеге асыру үшін getVarsforTerms негізгі әдісімен VariantRecognizer класы құрылды.

Осы бөлімді қорытындылай келе, алу рәсімінің әрқайсысы мәтінде терминдерді тұтынудың белгілі бір түрін тануға және өңдеуге мүмкіндік беретінін атап өтеміз. Сонымен қатар, рәсімдер арқылы алынған үміткер-терминдер жиынтықтары көбінесе бір-бірімен қиысады, өйткені бірдей сөз/сөз тіркесі әртүрлі рәсімдермен анықталуы мүмкін. Осылайша, белгілі бір ҒТ мәтінінен терминдерді автоматты түрде алу мәселесін шешу үшін осы рәсімдердің нәтижелерін біріктіру қажет, мұндай біріктіру стратегиясын біз олардың бөлек жұмысын эксперименттік зерттеу негізінде жасадық.

**4 АЛУ РӘСІМДЕРІН БІРІКТІРУ СТРАТЕГИЯСЫ**

4.1 Рәсімдердің жұмысын эксперименттік зерттеу

Терминологиялық ақпаратты алудың әзірленген рәсімдерін зерттеу екі пәндік саладан - информатика және есептеу техникасы (ИмЕТ) және физика ғылыми-техникалық мәтіндер жинағында жүргізілді; мәтіндердің жалпы көлемі 260 Кбайтты құрайды.

Жинаққа 7 мәтін кірді:

1-мәтін - «CASE-технологиялар. Ақпараттық жүйелерді жобалаудың заманауи әдістері мен құралдары» дәрістердің үзіндісі.

2-мәтін - жалпы салыстырмалылық теориясы туралы Википедия материалы [70].

3-мәтін - «Бағдарламалық өнімдер мен жүйелер» журналынан алынған «С++ тіліндегі разрядтылығы шексіз бүтін санды арифметикаға арналған cbignumber класы» мақаласы [71].

4-мәтін - «Микропроцессорлар» тақырыбындағы реферат [72].

5 - мәтін - «Күн оқиғалары» тақырыбы бойынша әр жылдардағы мақалалар (атап айтқанда [73]).

6-мәтін- «Эксперименттік және теориялық физика журналынан» алынған «Көп деңгейлі атомдарды Раман импульстарының бір фотонды шегінен төмен салқындату» мақаласы [74].

7-мәтін- «Функционалдық бағдарламалауды оқыту жүйесіндегі проблемалық-бағдарланған білім» диссертациясының шолу тарауы [75].

Рәсімдерді эксперименттік зерттеуде қолданылатын лексикалық-синтаксистік қалыптар базасының жалпы көлемі шамамен 6500 қалыпты құрады.

Зерттеуге getVarsforTerms рәсімінен басқа барлық әзірленген алу рәсімдері қатысты: бұл рәсім басқа рәсімдер арқылы алған үміткер-терминдердің жиынтықтарымен жұмыс істейді және осыған қарамастан олардың жұмысын бағалаудың мағынасы жоқ.

Рәсімдердің нәтижелерін бағалау алынған үміткер-терминдерінің жиынтығын сарапшылар құрастырған үлгілік терминдер жиынтығымен салыстыру арқылы жүргізілді. Үлгілік жиынтықтарды қалыптастыру барысында терминологиялық сөздіктердің толық еместігі анық байқалғанын ескертеміз [76]: мысалы, *ақпараттық жүйе* тіркесі ИмЕТ сөздігінде болған, ал оның жалпы қабылданған *АЖ* қысқартылуы - жоқ.

Рәсімдердің нәтижелерін бағалау үшін талданатын мәтіндегі терминдерді алудың толықтығы мен дәлдігі есептелді. Терминдерді алудың толықтығы деп әр түрлі NR терминдерінің рәсімі бойынша дұрыс табылған санның мәтіндегі олардың жалпы санына қатынасы түсініледі:

Rt= ( NR/ NT) \* 100 %

Pt терминдерін алудың дәлдігі деп дұрыс табылған NR терминдер рәсімдері санның NF терминдерге үміткерлердің рәсім бойынша табылған жалпы санына қатынасы түсініледі:

Pt= ( NR/ NF) \* 100 %

Сонымен қатар, getDictTerms, getAuthTerms және getNonDictTerms рәсімдері үшін олар таныған терминдердің әрқайсысының әр түрлі қолданыстарын алудың толықтығы мен дәлдігі есептелді, бұл мәтіндегі терминдердің тұтынылу жиілігін анықтаудың дұрыстығын тексеруді білдіреді. Ru толықтығы мен Рu дәлдігі қолданысын алу Rt және Рt  -ге ұқсас формулалар бойынша есептелді, онда NR деп рәсімдері бойынша дұрыс табылып танылған терминдердің әрқайсысының әртүрлі қолдану саны түсініледі, Nт мәтінде танылған терминдердің әрқайсысының жалпы қолданылу саны, ал NF деп терминдерді қолданудың рәсімі бойынша табылған жалпы саны түсінілді.

4.1-кестеде алу рәсімінің толықтығы мен дәлдігінің мәндері келтірілген: Rt және Pt - әртүрлі терминдерді алудың толықтығы мен дәлдігі, Ru және Pu - олардың тұтынуын анықтаудың толықтығы мен дәлдігі. Толықтығы мен дәлдігі сарапшылар жасаған үлгілік жиынтықтарға қатысты анықталды.

Кесте 4.1. Терминдер мен оларды тұтынудың толықтығы мен дәлдігі

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Рәсім атауы | Терминдер | | Тұтынылуы | |
| Rt | Pt | Ru | Pu |
| getDictTerms | 94,0% | 83,2% | 89,2% | 72,0% |
| getNonDictTerms | 57,7% | 27,4% | 59,6% | 48,6% |
| getAuthTerms | 92,3% | 95,9% | 73,7% | 77,9% |
| getSynTerms | 64,0% | 49,9% | — | — |
| getTermsfrmCombs | 81,7% | 24,7% | — | — |

4.1-кестеден getDictTerms сөздіктік терминдерінің кіру рәсімін алу Rt, Pt және Ru–да ең жоғары мәндерді көрсеткенін байқауға болады, бұл ондағы терминологиялық сөздіктің қолданылуымен түсіндіріледі. Дәлдіктің ең жоғары Pu мәнін getAuthTerms авторлық терминдерді шығару рәсімі көрсетті, онда терминдерді тану олардың анықтамалық қалыптарына негізделеді.

Тұтастай алғанда, алу рәсімінің жұмысын зерттеу кезінде алынған үміткер-терминдер жиынтықтары бір-бірімен қиылысады, толық емес және терминдік емес сөздер мен сөз тіркестері көп мөлшерді қамтиды, бұл оларды әрі қарай өңдеуді қажет етеді.

# 4.2 Зерттеу нәтижелерін талдау

Үміткер-терминдердің жиынтығын терминдердің үлгілік жиынтығымен салыстыру терминдерді алу мен тұтынудың толықтығы мен дәлдігінің жоғары мәндері жеткіліксіздігінің негізгі себептерін анықтады.

Қалып базасының толық еместігі

Барлық зерттелетін рәсімдердің нәтижелеріне терминді тұтынудың сирек жағдайлары ескерілмеген қалыптар базасының толық болмауы әсер етті; шығындар орташа есеппен Rt-ге - 14,9%, Pt үшін - 1,7%, Ru үшін - 5,4% құрады. Мысалы, 4-мәтінде мына сөйлемше кездесті

*Регистр - бұл бистабильді құрылғылардың жиынтығын білдіреді* кіретін авторлық термин *регистр*, алайда оны танитын лексикалық-синтаксистік қалып базада жоқ. Қалыптар базасына осы құралым қалыбын қосқанда, авторлық терминдерді алудың толықтығы 5,7%- ға артады, бірақ дәлдігі 25,3%- ға төмендейді. Дәлдіктің төмендеуі сөзсіз осы құралымның түсініксіздігінен туындайды, ол түсініксіздік *білдіреді* сөзіне негізделген және мәтіннің анықтамасын енгізудің орнына көбінесе қандай да бір ұғымды түсіндіреді.

Лексикалық-синтаксистік қалыптардың ерекшеліктері

Көбінесе бірдей тілдік құралымдарға әртүрлі қалыптар қолданылатын жағдай туындайды, көптеген жағдайларда бұл терминдер мен оларды тұтынудың толықтығы мен дәлдігінің нашарлауына әкеледі. Көрсетілген себептерге байланысты толықтығының жоғалуы Rt үшін 28,8% және Ru үшін 30,1%, дәлдіктің жоғалуы Rt үшін 3,0% және Рu үшін 13,1% құрады.

Мысалы, 1-мәтінде N1 А2 N2<с=gen><А2=N2> қалыбы бойынша getNonDictTerms рәсімі термин болып табылмайтын *динамикалық жадыны бөлу* тіркесін анықтады, ал А1 N1 <А1=N1> қалыбымен сипатталған *динамикалық жад* термині анықталған жоқ, өйткені ол оның бөлігі болып шықты. Сонымен қатар, 3-мәтінде getDictTerms рәсімі *анықтама* сөздіктік терминін анықтады, ол бұл мәтінде *санның қарапайымдылығын анықтау* сөздіктік емес терминнің бір бөлігі ғана болады.

Бұдан басқа, кейбір тілдік құралымдар ұқсас синтаксистік құрылымға ие, мысалы, жеке терминдер және бірнеше терминологиялық тіркестердің шылаусыз қосылыстары (терминдерді қараңыз: *тәуелсіз алгоритмдер әдісі*, *үлкен разрядтылық саны* және қосылыстар - *динамикалық жадыны бөлу*, *сыртқы деректер желісі*). Осылайша, 1-мәтіндегі *динамикалық жадыны бөлу* тіркесі шын мәнінде *жадыны бөлу* сөздіктік термині және *динамикалық жады* сөздіктік емес терминінің қосындысы болып табылады.

Демек, барлық зерттелетін рәсімдер бойынша алынған үміткер-терминдер жиынтықтары жоғарыда сипатталған терминдерді кірістіру және біріктіру жағдайларын анықтау үшін жекеленген бірлескен өңдеуді қажет етеді.

Лингвистикалық критерийлердің шектеулері

Бұл себеп барлық алу рәсімдерінде дәлдіктің айтарлықтай төмендеуіне әкеледі, шығындар орташа есеппен Pt үшін - 36,8%, Рu үшін - 15,6% құрады. Мысалы, мәтіндегі терминдердің типтік синтаксистік үлгілері терминдер болып табылмайтын немесе басқа рәсімдер арқылы терминдер ретінде анықталған көптеген сөздер мен сөз тіркестеріне сәйкес келеді. 3-мәтіннен А N <А=N> үлгісіне сәйкес келетін, бір жағынан, *әртүрлі тәсілдер* мен *қарапайым шешім* сияқты тіркестер табылса, екінші жағынан, *жасырын қате* және *ұзын сан* секілді сөздіктік емес терминдер анықталды.

Осылайша, алу рәсімдері арқылы алынған үміткер-терминдер жиынтығын олардан термин болып табылмайтын сөздер мен сөз тіркестерін алып тастау үшін қосымша өңдеу керек. Мысалы, терминдерді алу дәлдігін жақсарту үшін getNonDictTerms рәсімінің нәтижелерін өңдеу кезінде ең аз ықтимал терминдерді електен өткізуге мүмкіндік беретін статистикалық критерийді тартқан жөн. Шекті мән ретінде үміткер-терминдердің барлық жиіліктерінің өлшенген арифметикалық орташа мәнін қолдану ұсынылады (осы рәсім арқылы алынған), бұл олардың маңызды сипаттамаларын ескереді:

мұндағы f1– осындағы i-oro үміткердің тұтыну жиілігінің мәні, ал бөлшек бөлгіш - бұл getNonDictTerms рәсімімен анықталған сөздер мен сөз тіркестері саны. getNonDictTerms рәсімімен алынған барлық үміткер-терминдерді қарастырған кезде, алу дәлдігі 27,4% құрайды, ал тек тұтыну жиілігі F дөңгелектелген мәннен кем емес кандидаттарды қарастырғанда дәлдігі қазірдің өзінде 71,5% құрайды.

Мәтіндік нұсқаларды тұтыну

getDictTerms, getAuthTerms және getNonDictTerms рәсімдерінде терминдерді алу және оларды тұтыну толықтығының төмендеуі сонымен қатар олардың (әрқайсысында жеке) терминдерді түрлендіру мүмкіндігін ескермейтіндігімен байланысты; орташа шығын Rt үшін - 5,7%, Ru үшін - 7,8%, Pt үшін - 4,0%, Рu үшін - 3,4% құрады. Мысалы, 2-мәтіндегі *жалпы салыстырмалылық теориясы* терминін қолдану жиілігі сарапшылардың бағалауы бойынша 37-ге тең, оның ішінде 15 рет термин мәтінде көрсетілген түрде қолданылады, оның *салыстырмалылық теориясы* деп берілетін қиылған нұсқасы 1 рет кездеседі, ал қысқартылған-синонимі - *ЖСТ* мәтінде 21 рет қолданылады. getDictTerms рәсімі осы терминнің тек 15 қолданысын тапты.

Сөздіктік терминдерді алу кезінде синонимдерді есепке алу (getDictTerms және getSynTerms рәсімдердің жұмыс нәтижелерін біріктіру арқылы) бір терминнің әртүрлі қолданыстарын алудың толықтығын 3,9%-ға арттыруға мүмкіндік беретінін ескереміз. Екінші жағынан, рәсімдердің нәтижелерін осылайша біріктіру алу дәлдігі мен терминдердің және оларды қолдану дәлдігін сәйкесінше 7,8% және 6,0% төмендеуіне әкеледі. Демек, синоним-терминдерді, қосылыстардағы терминдерді және мәтіндік нұсқаларды (getSynTerms, getTermsfromCombs және getVarsforTerms рәсімдерінің жұмыс нәтижелерін сәйкесінше) тұтыну нұсқаларын есепке алудың белгілі бір стратегиясы қажет, бұл тұтастай алғанда алудың толықтығы мен дәлдігін жақсартуға мүмкіндік береді.

# 4.3 Рәсімдердің нәтижелерін біріктіру стратегиясы

Ұсынылған стратегияның мақсаты - терминдердің ҒТ мәтінінен алу және оларды тұтынудың толықтығы мен дәлдігін жақсарту; оның мәні келесідей [77]. Алдымен, терминдерді тұтынудың өңделген алу рәсімдері қарастырылып отырған мәтінге жеке қолданылады, оларды қолдану нәтижесінде үміткер-терминдер жиынтығы алынады. Содан кейін олардың ішінен эксперименттік зерттеу нәтижелері бойынша тұжырымдалған эвристикалық ережелер бойынша ең ықтимал үміткерлер таңдалады. Ережелер кезек-кезек қолданылады, нәтижесінде өңделген мәтін үшін таңдалған үміткер-терминдер туралы М жазбалардың жиынтығы және G1,G2,..., GN мәтіндік нұсқалар топтарының қорытынды жинағы итеративті түрде құрылады.

Ережелер үш топқа бөлінеді:

1) М жиынтығын бастапқы қалыптастырудың 1-3 ережесі.

2) терминдерді тұтыну нұсқаларын ескере отырып, М жиынтығын кеңейтудің 4-7 және 9-12 ережелері.

3) G1,G2,...,GN. мәтіндік нұсқаларының топтарын құрудың 8-ережесі.

Ережелер үміткер-терминдерді М жиынтығына және G1, G2,..., GN топтарының жинағына қосу шарттарын белгілейді. Ережелерді қолдануды бастамас бұрын [F] мәні есептеледі - Мnondict жиынтығының барлық элементтері (getNonDictTerms рәсімі бойынша алынған) жиіліктерінің дөңгелектелген орташа өлшенген арифметикалық мәні. Ережелерді нөмірлеу оларды қолдану тәртібіне сәйкес келеді. Үміткерді М жиынтығына қосқан сайын, getFrequency рәсімі шақырылады.

Біз келесі мәтіндік үзінді мысалында ережелердің қолданылуын суреттейміз:

*Микропроцессор - бұл бірыңғай жартылай өткізгіш кристалда жүзеге асырылатын және орталық процессордың функцияларын орындауға қабілетті аса үлкен интегралды схема. Сыртқы құрылғылармен (СҚ) микропроцессор микросхема корпусының арнайы түйреуіштеріне шығарылған мекенжай, деректер және басқару құрсымдарының арқасында «сөйлесе» алады. Бұл жағдайда микропроцессордың ішкі регистрлерінің разрядтылығы деректер желілері үшін сыртқы түйреуіштер санына сәйкес келмеуі мүмкін, яғни 32 разрядтық регистрлері бар микропроцессорда, мысалы, тек 16 сыртқы деректер желісі болуы мүмкін.*

*Микропроцессорға қатысты енгізу-шығару операцияларын жасайтын кез келген сыртқы құрылғыны перифериялық деп атауға болады. Порт - бұл әдетте бір немесе бірнеше енгізу-шығару регистрлерін қамтитын және, мысалы, перифериялық құрылғыны микропроцессордың сыртқы құрсымдарына қосуға мүмкіндік беретін жұптастыру схемасы. Әрбір дерлік микросхема әртүрлі мақсаттарда бірнеше енгізу-шығару порттарын пайдаланады.*

М жиынтығының бастапқы қалыптастыру ережелері

М жиынтығының бастапқы қалыптасу ережелерінде келесі жиынтықтар қолданылады:

* getAuthTerms рәсімінен алынған Мauth авторлық үміткер-терминдерінің жиынтығы;
* getDictTerms рәсімінен алынған Мdict сөздіктік үміткер-терминдерінің жиынтығы;
* getNonDictTerms рәсімі бойынша алынған Мnondict сөздіктік емес үміткер-терминдерінің жиынтығы;
* T мәтінге шылаусыз қосылыстардың қалыптарын қабаттастыру арқылы алынған Мсоmb шылаусыз қосылыстарының жиынтығы; әрбір қосылыстан getTermsfromCombs рәсімі арқылы оның құрамдас үміткер-терминдері анықталады.

Қарастырылып отырған мысал үзіндісінен алынған үміткер-терминдерінің жиынтығы 4.2-кестеде келтірілген. Қалың әріппен бөлектелген:

* Мdict жиынтығындағы үміткерлер - Мauth жиынтығында да бар;
* Мnondict- ке үміткерлер - Мdict-ке де қатысатын үміткерлер;
* Мсоmb-ға үміткерлер - Мauth және Мdict-ке қатысатын үміткерлер;

Сонымен қатар, Мnondict -те Мсоmb. элементтері сияқты синтаксистік үлгілермен сипатталған үміткерлерге баса назар аударады. Қарастырылып отырған үзінді үшін [F] мәні 1-ге тең.

Кесте 4.2. Үміткер-терминдер мен қосылыстар жиынтығы

|  |  |
| --- | --- |
| Жиынтық | Жиынтық құрамы |
| Мauth- авторлық терминдер | *порт, перифериялық құрылғы* |
| Мdict-сөздіктік терминдер | *шығару, енгізу-шығару, деректер, желі, микропроцессор, операция,* ***порт*,** *енгізу-шығару порты, разрядтылық, регистр, жұптастыру, орталық процессор, құрсым, мекенжай құрсымы* |
| Мnondict –сөздіктік емес терминдер | *32-разрядтық регистр, микропроцессордың сыртқы құрсымы,сыртқы құрылғы, СҚ, процессордың ішкі регистрі,* ***деректер****, бірыңғай жартылай өткізгіш кристалл, сыртқы шығару саны, микросхема корпусының түйіспесі, сыртқы деректер желісі, деректер желісі,* ***микропроцессор***, *енгізу-шығару операциясының микропроцессоры, микросхема, қатынас,* ***перифериялық құрылғы****,* ***порт*, *енгізу-шығару порты***, *түрлі мақсаттар, разрядтылық ішкі регистрларды, енгізу-шығару регистрі, аса үлкен интегралды схема, корпустың арнайы түйреуіші, жұптастыру схемасы, басқару, орталық процессор функциясы,* ***мекенжай құрсымы*** |
| Мсоmb –шылаусыз қосылыстар | *аса үлкен интегралды схема, бірыңғай жартылай өткізгіш, микропроцессордің ішкі регистр разрядтылығы, орталық процессор функциясы, ішкі регистр разрядтылығы, сыртқы шығару саны, сыртқы деректер желісі, микропроцессордың сыртқы құрсымы, микропроцессордың ішкі регистрі, микросхема корпусының түйіспесі, деректер желісі, енгізу-шығару операциясының микропроцессоры , енгізу-шығару порты,* ***енгізу-шығару*** *регистрі, корпустың арнайы түйреуіші,* ***мекенжай құрсымы****, жұптастыру схемасы* |

Авторлық терминдер салыстырмалы түрде толық және дәл шығарылады, сондықтан:

1-Ереже. Мauth-тың барлық үміткер-терминдері М жиынтығына енгізілген.

Қарастырылып отырған мысал үзіндісі үшін М жиынтығына *порт* және *перифериялық құрылғы* үміткер-терминдері енгізіледі.

Сөздіктік терминдер де барынша толық және дәл анықталады, бірақ лексикалық-синтаксистік қалыптардың ерекшеліктерін ескере отырып, ссөздіктік терминдер басқа терминдердің құрамдас бөлігі болып табылатын немесе болмайтын жағдайларды ажырату қажет. Мdict жиынтығының үміткер- термині Мnondict-дің бөлігі болып табылмайтын жағдайларды өңдеу үшін 2-ереже қолданылады, керісінше жағдайда 3-ережені пайдаланамыз.

2-ереже. Мdict жиынтығындағы үміткер-термині, егер ол Мnondict-тегі үміткер-терминінің бөлігі болмаса, М жиынтығына қосылады. 2-ереже бойынша М жиынтығына келесі сөздіктік үміткер-терминдер қосылады: *енгізу-шығару, деректер, сызық, микропроцессор, енгізу-шығару порты, регистр, мекенжай құрсымы* (*порт* термині 1-Ережеге сәйкес М-ге енгізілген). Басқа Мdict үміткер-терминдері Мnondict элементтердің құрамдас бөліктері болып табылады және М-ге енгізілмеген (мысалы, *орталық процессор* сөздіктік термині Mnondict–тегі *орталық процессор функциясы* терминіне кіреді).

3-ереже. Егер Mnondict-тегі term үміткер-термині сонымен қатар Mcomb жиынтығына кірсе, яғни терминдердің қосындысы болуы мүмкін және term терминді құрайтын үміткерлер Mdict жиынтығының элементтері болып табылса, онда бұл кандидаттардың барлығы M жиынтығына қосылады және қарастырылып отырған term үміткер-термин Mnondict және Mcomb жиынтығынан алынып тасталады.

Әйтпесе, егер қарастырылып отырған term терминнің құрамдас бөліктері арасында Mdict жиынтығынан тек бір үміткер болса және ол тек осы қосылыс шеңберінде мәтінде қолданылса, онда Mnondict-тің term терминнің өзі M жиынтығына қосылады. Сипатталған әрекеттерден кейін Mnondict жиынтығынан М жиынтығына қосылған барлық элементтер жойылады.

3-ереже бойынша М жиынтығына *операция* сөздіктік үміткер-термині қосылады, өйткені Mnondict және Mcomb жиынтықтарына кіретін *енгізу-шығару операциялары микропроцессоры* тіркесімі үш сөздіктік терминдердің шылаусыз қосындысы болады: *микропроцессор, операция* және *енгізу-шығару*.

Сондай-ақ, 3-ережеге сәйкес М жинақтығына Mnondict-тен *жұптастыру схемасы* тіркесі енгізіледі: ол Mcomb жиынтыққа кіреді, бірақ оның құрамына бір сөздіктік термин кіреді - *жұптастыру* (*схема* сөзі Mdict жиынтығына кірмейді), ол тек қарастырылып отырған сөз тіркесінің бөлігі ретінде қолданылады. *Жұптастыру* сөздіктік терминінің өзі М жиынтығына қосылмайды.

4.3-кестеде 1-3 ережелерін қолдану нәтижесінде алынған М жиынтығы және өзгертілген үміткер-терминдерінің жиынтығы берілген: М жиынтығына кірген үміткерлер тиісті жиынтықтардан сызылып тасталған. Біздің мәтінге арналған сөздіктік терминдерден мысал ретінде М жиынтығына тек *құрсым* термині енгізілмегенін ескереміз.

Кесте 4.3. 1-3 ережелерді қолданғаннан кейінгі үміткерлер жиынтығы

|  |  |
| --- | --- |
| Жиынтық | Жиынтық құрамы |
| М – алынған үміткерлер | порт, перифериялық құрылғы (1-ереже), енгізу-шығару, деректер, желі, микропроцессор, енгізу-шығару порты, регистр, мекенжай құрсымы (2-ереже) шығару, разрядтылық, орталық процессор, операция, схема жұптастыру (3-ереже) |
| Мauth - авторлық терминдер | *~~порт, перифериялық құрылғы~~* |
| Мdict- сөздіктік терминдер | ~~шығару, енгізу-шығару, деректер, желі, микропроцессор, операция, енгізу-шығару порты, разрядтылық, регистр, жұптастыру, орталық процессор,~~ құрсым~~, мекенжай құрсымы~~ |
| Мnondict- сөздіктік емес терминдер | 32-разрядтық регистр, микропроцессордың сыртқы құрсымы,сыртқы құрылғы, СҚ, микропроцессордың ішкі регистрі, бірыңғай жартылай өткізгіш, сыртқы шығару саны, микросхема корпусының түйіспесі, сыртқы деректер желісі, деректер желісі, енгізу-шығару операциясының микропроцессоры, микросхема, қатынас , түрлі мақсаттар, разрядтылық ішкі регистрларды, енгізу-шығару регистрі, аса үлкен интегралды схема, корпустың арнайы түйреуіші, ~~жұптастыру схемасы~~, басқару, орталық процессор функциясы |
| Мcomb – шылаусыз қосылыс | аса үлкен интегралды схема, бірыңғай жартылай өткізгіш кристалл, микропроцессордың ішкі регистр разрядтылығы, орталық процессор функциясы, ішкі регистр разрядтылығы, сыртқы шығару саны, сыртқы деректер желісі, микропроцессордың сыртқы құрсымы, микропроцессордың ішкі регистрі, микросхема корпусының түйіспесі, деректер желісі, енгізу-шығару операциясының микропроцессоры, ~~енгізу-шығару порты~~, енгізу-шығару регистрі, корпустың арнайы түйреуіші, ~~мекенжай құрсымы~~, жұптастыру схемасы |

М жиынтығының кеңейту ережелері

М жиынтығының кеңейту ережелерінде мына жиынтықтар қолданылады:

* таңдалған үміткер-терминдерінің ағымдағы М жиынтығы;
* 1-3 ережелерін қолданғаннан кейін өзгертілген Мdict, Мnondict және Мcomb жиынтықтары;
* getSynTerms рәсімі арқылы алынған <термин-синоним> Мsyn жұптар жиынтығы;
* Т мәтінге шылаулы қосылыстардың қалыптарын қабаттасу арқылы алынған Мconj шылаулы қосылыстар жиынтығы; getTermsfromCombs рәсімі арқылы қосылыстардың әрқайсысы үшін оның құрамдас үміткер- терминдері анықталған.

М, Mdict, Mnondict және Мcomb жиынтықтары 4.3-кестеде көрсетілген. Біздің мысал мәтінінде Мsyn жиынтығына арналған тек бір жұп бар: *сыртқы құрылғы*-*СҚ*. Мconj жиынтығына *мекенжай құрсымы, деректер* және *басқару* қосылымы кіреді, одан getTermsfromCombs рәсімімен *мекенжай құрсымы*, *деректер құрсымы* және *басқару құрсымы* терминдерін алады.

Біздің стратегиямыздың 4-6 ережелері терминдерді тұтыну нұсқаларын ескеруге мүмкіндік береді. Синонимдерді өңдеу үшін қолданылады:

4 ереже. Егер Мsyn жиынтығының кейбір <термин-синоним> жұбы үшін оның бір элементі М жиынтығына енсе, онда екіншісі де М жиынтығына қосылуы керек.

Қарастырылып отырған Мsyn үшін үзінді-мысал *сыртқы құрылғы - СҚ* жұбын қамтиды, бірақ *сыртқы құрылғы* термині де, *СҚ* термині де әлі М жиынтығына кірмегендіктен, 4 ереже қолданылмайды.

Келесі екі ереже сәйкесінше шылаулы және шылаусыз қосылыстарды өңдеуге қызмет етеді.

5 ереже. Егер шылаулы қосылыстың құрамына (Мconj жиынтығының құралымы) кем дегенде бір элемент кірсе:

* М жиынтығынан, немесе
* Mdict жиынтығынан, немесе
* Mnondict жиынтығынан үзілген түрде, онда қосылысты құрайтын барлық терминдер М жиынтығына қосылады. Қарастырылған қосылыс Мconj жиынтығынан жойылады.

Мconj-де *мекенжай*, *деректер және басқару құрсымы* шылаулы қосылыс бар. *Мекенжай құрсымы* термині қазірдің өзінде М жиынтығына енгендіктен, содан кейін 5-ережеге сәйкес М жиынтығына көрсетілген қосылымды құрайтын терминдер енгізіледі: *деректер құрсымы* және *басқару құрсымы* (*мекенжай құрсымы* термині М жиынтығында бар болғандықтан оған қайта қосылмайды).

6 ереже. Егер шылаусыз қосылыстың құрамына (Мcomb жиынтығының элементі) Мdict, М немесе Мnondict жиынтықтарының кем дегенде бір элементі үзілген түрде кірсе, онда бұл қосылысты құрайтын барлық элементтер М жиынтығына қосылады және тиісті жиынтықтардан жойылады.

Егер мұндай элементтер болмаса, онда біз қосылыстың өзін термин деп санаймыз және оны М жиынтығына қосамыз. Қарастырылған қосылыс Мcomb және Мnondict жиынтықтарынан жойылады.

6-ережеге сәйкес, мысалы, *сыртқы деректер желісі* қосылысы өңделеді, ол сөздіктік емес үміткер-терминінен тұрады *деректер желісі* (оның кіруі үзілген) және *сыртқы деректер* тіркестері; екі тіркес те М жиынтығына қосылады. Сол ережеге сәйкес Мcomb жиынтығының *аса үлкен интегралды схема* және *бірыңғай жартылай өткізгіш кристалл* қосылыстары толығымен М жиынтығына қосылады.

7 ереже. Егер 4-6 ережелеріне сүйене отырып, М жиынтығына жаңа элементтер енгізілген болса (М жиынтығы кеңейген), онда бұл ережелерді қайта қолдану қажет.

Қарастырылып отырған мысал үзіндісі үшін 4-6 ережелерін қайта қолдану (7-ережеге сәйкес) М жиынтығын кеңейте бермейтіндіктен, 7-ереже тек бір рет қолданылады.

Мысал үзінді үшін 4-7 ережелерін қолдану нәтижесі 4.4-кестеде келтірілген. Mnondict, Мcomb және Mconj жиынтықтарында өңделген элементтер сызылған, ал М жиынтығында ұзынырақ (асты сызылған) терминдердің бөліктері болып табылатын үміткер-терминдердің асты сызылған және осы себепті жиынтықтан үміткер сөзді шығарған сайын шақырылатын getFrequency рәсімінің жұмысы нәтижесінде М жиынтығынан жойылған.

Кесте 4.4. 4-7 ережелерін қолданғаннан кейінгі үміткерлер жиынтығы

|  |  |
| --- | --- |
| Жиынтық | Жиынтық құрамы |
| М-алынған үміткерлер | *порт, перифериялық құрылғы* (1-ереже)  *енгізу-шығару, деректер, желі, микропроцессор, енгізу-шығару порты, регистр, мекенжай құрсымы* (2-ереже)  *шығару, разрядтылық, орталық процессор, операция, схемасы*  *жұптастыру* (3-ереже)  *деректер құрсымы, басқару құрсымы* (5ереже)  *аса үлкен интегралды схема, бірыңғай жартылай өткізгіш кристалл, ішкі регистрі, регистр разрядтылығы, деректер желісі, сыртқы деректер, сыртқы құрсым* (6-ереже) |
| Мdict-сөздіктік терминдер | *құрсым* |
| Мnondict-сөздіктік емес терминдер | *32-разрядтық регистр, микропроцессордың сыртқы құрсымы, сыртқы құрылғы, СҚ, микропроцессордың ішкі регистрі, бірыңғай жартылай өткізгіш кристалл, сыртқы шығару саны, микросхема корпусының түйіспесі, сыртқы деректер желісі, деректер желісі, енгізу-шығару операциясы микропроцессоры, микросхема, қатынас, түрлі мақсаттар, ішкі регистр разрядтылығы, енгізу-шығару регистрі, аса үлкен интегралды схема, корпустың арнайы түйреуіші, жұптастыру схемасы, басқару, орталық процессор функциясы* |
| Мcomb-шылаусыз қосылыстар | *аса үлкен интегралды схема, бірыңғай жартылай өткізгіш кристалл, микропроцессордың ішкі регистр разрядтылығы, орталық процессор функциясы, ішкі регистр разрядтылығы, сыртқы шығару саны, сыртқы деректер желісі, микропроцессордың сыртқы құрсымы, микропроцессордың ішкі регистрі, микросхема корпусының түйіспесі, деректер желісі, енгізу-шығару операциясы микропроцессоры, енгізу-шығару регистрі, корпустың арнайы түйреуіші, мекенжай құрсымы, жұптастыру схемасы* |
| Msyn- синонимдер | *сыртқы құрылғы – СҚ* |
| Mconj- шылаулы қосылыстар | *Мекенжай, деректер және басқару құрсымы* |

G1, …, GN мәтіндік нұсқаларының топ жинақтарын құру ережесі

8-ереже терминдердің дублеттері мен лексикалық-синтаксистік нұсқаларын есепке алу үшін қолданылады - онда 4-7 ережелерін (4.4 кестені қараңыз) қолданғаннан кейін алынған М және Mnondict  жиынтықтары пайдаланылады. Нұсқаларды іздеуге getVarsfromTerms рәсімі шақырылады, оның шығуында сөздер мен сөз тіркестерінің G1, G2, …, GN топтары бір терминнің нұсқалары болып саналады.

8-ереже. М жиынтықтарынан үміткер-терминдердің дублеттерін іздеу М жиынтығының элементтері арасында жүзеге асырылады. М жиынтығынан үміткер- терминдерінің лексикалық-синтаксистік нұсқаларын іздеу М және Мnondict жиынтықтарының элементтері арасында жүзеге асырылады. Табылған нұсқалар Мnondict жиынтығынан жойылады.

Осы Ережеге сәйкес, біздің мысалда М жиынтығындағы сөз тіркестерінің ішінен лексикалық-синтаксистік нұсқалардың келесі топтары табылды: G1= {енгізу-шығару порты, порт}; G2= {ішкі регистр, регистр}. Мnondict жиынтығының элементтері арасынан нұсқалар табылмады.

М жиынтығын одан әрі кеңейту ережелері

Келесі ереже 8-ереже бойынша анықталған мәтіндік нұсқалардың синонимдерінің болуы және басқа терминдермен байланысқа түсуі мүмкін екенін ескеруге арналған. 9-ережеде 4-8-ережелерді қолдану нәтижесінде алынған жиынтықтар қолданылады.

9-ереже. Егер 8-ереже бойынша анықталған мәтіндік нұсқаның ішінде М жиынтығына кірмейтіндер болса, онда олар оған қосылады. Алынған М жиынтығы үшін 4-8 ережелері қайта қолданылады.

Қарастырылып отырған үзінді-мысал үшін 9-ереже М жиынтығын кеңейтпейді, өйткені 8-ережені қолдану нәтижесінде *порт* пен *регистр* мәтіндік нұсқалары алынды, олар қазірдің өзінде М жиынтығында бар.

Стратегияның келесі ережелерінде (10, 11 және 12 - ережелер) 9-ережені қолдану нәтижесінде алынған М, Мdict, Мnondict және Мcomb, жиынтықтары қолданылады (4-8-ережелерді қайталау) - 4.5-кестені қараңыз.

4.5-кесте. 9-ережені қолданғаннан кейінгі үміткерлер жиынтықтары

|  |  |
| --- | --- |
| Жиынтық | Жиынтық құрамы |
| М – алынған үміткерлер | *порт, перифериялық құрылғы* (1-ереже) *енгізу-шығару, микропроцессор, енгізу-шығару порты, регистр, мекенжай құрсымы* (2-ереже)  *шығару, орталық процессор, операция, жұптастыру схемасы* (3-ереже)  *деректер құрсымы, басқару құрсымы* (5-ереже) *аса үлкен интегралды схема, бірыңғай жартылай өткізгіш кристалл, ішкі регистр, регистрлар разрядтылығы, деректер желісі, сыртқы деректер, сыртқы құрсым* (6-ереже) |
| Мdict - сөздіктік терминдер | *құрсым* |
| Мnondict – сөздіктік емес терминдер | *32-разрядты регистр, сыртқы құрылғы, СҚ, сыртқы шығару саны, микросхема корпусының түйіспесі, микросхема, қатынас, түрлі мақсаттар, корпустың арнайы түйреуіші, орталық процессор функциясы* |
| Мсоmb – шылаусыз қосылыстар | *орталық процессор функциясы, сыртқы шығару саны, микросхема корпусының түйіспесі, корпустың арнайы түйреуіші* |

Келесі ереже Мnondict жиынтығында қалған терминдердің ішінен ең ықтимал үміткерлерді таңдауға көмектеседі.

10 ереже. Mnondict жиынтығының элементі, егер ол Mcomb жиынтығына кірмесе және оны қолдану жиілігі [F] - дөңгелектелген орташа өлшенген арифметикалық жиіліктерден кем болмаса М-ге қосылады. Қосылған элемент Mnondict- тен жойылады.

10-ереже бойынша М жиынтығына *32 разрядтық регистр*, *сыртқы құрылғы, СҚ, микросхема, қатынас, әртүрлі мақсаттар* терминдері қосылады.

М жиынтығына жаңа қосылған үміткер-терминдерді тұтыну нұсқаларын есепке алу үшін 11-ереже қолданылады; онда 9 және 10 ережелерін қолдану нәтижесінде алынған жиынтықтар пайдаланады.

11 ереже. Егер 10-ережені қолдану нәтижесінде М жиынтығы кеңейсе, онда оған 4-9 ережелері қайта қолданылады.

Осы ережені қолдану нәтижесінде (8-ереже бойынша) лексикалық-синтаксистік нұсқаларының *32 разрядтық регистр - регистр* жұбы табылатын болады, бірақ олардың екеуі де М жиынтығына енгізіліп қойған.

Стратегияның соңғы ережесі өңделмеген, яғни осы уақытқа дейін Mdict–те қалған сөздіктік терминдерді есепке алуға қызмет етеді; онда 11-ережені қолдану нәтижесінде алынған жиынтықтарды пайдаланады.

12 ереже. Егер М жиынтығына Mdict-тегі үміткер-терминдер қосқаннан кейін М жиынтығы кеңейсе, ол үшін 4-9 ережелерін қайталаймыз.

Қарастырылып отырған үзінді-мысал үшін Mdict жиынтығындағы терминдер М-ге қосылса, бұдан М жиынтығы кеңеймейді (*құрсым* сөздіктік термині М-дегі *сыртқы құрсым* үміткер-терминінің бөлігі болып табылады, мұны getFrequency рәсімі анықтайды).

Осымен стратегияны үзінді-мысалға қолдану аяқталады. Барлық ережелерді қолдану нәтижесінде алынған іріктелген М үміткерлерінің қорытынды жиынтығы және сарапшылар осы үзінді үшін құрастырған Мехр терминдерінің үлгілік жиынтығы 4.6-кестеде келтірілген. Жақшада әр термин оны қолдану жиілігі көрсетіледі, екі жиынтықта да бар терминдер асты сызылған. Барлық анықталған терминдердің стратегиясының нәтижесінде тұтыну жиілігі дұрыс есептелгені маңызды.

Кесте 4.6. Үзінді-мысалға арналған қорытынды және сараптамалық жиынтық

|  |  |
| --- | --- |
| Жиынтық | Жиынтық құрамы |
| M- алынған үміткерлер | *32-разрядтық регистр (1), енгізу-шығару (1),сыртқы құрылғы (1), СҚ (1), сыртқы деректер (1), сыртқы құрсымы (1), ішкі регистр (1), шығару (1), бірыңғай жартылай өткізгіш кристалл (1),деректер желісі (2), микропроцессор* (6), *микросхема* (2), *операция (1), қатынас (1), перифериялық құрылғы* (2), *порт (1), енгізу-шығару порты (1), түрлі мақсаттар (1), регистр разрядтылығы (1), аса үлкен интегралды схема (1), жұптастыру схемасы (1), орталық процессор (1), мекенжай құрсымы (1), деректер құрсымы (1), басқару құрсымы (1)* |
| Mexp – мәтін терминдері | *32-разрядтық регистр (1), сыртқы құрылғы (1), СҚ (1), сыртқы шығару (1), сыртқы құрсым (1), ішкі регистр (1), деректер желісі* (2), *микропроцессор (6), микросхема* (2), *енгізу операциясы (1), перифериялық құрылғы* (2), *порт (1), енгізу-шығару порты (1), енгізу-шығару регистрі (1), аса үлкен интегралды схема(1), жұптастыру схемасы (1), орталық процессор (1), мекенжай құрсымы (1), деректер құрсымы (1), басқару құрсымы (1)* |

М қорытынды жиынтығына сарапшылар анықтаған 20 терминнің тек 3-і кірген жоқ. Сөздіктік терминдердің қосындысы ретінде *енгізу-шығару операциясы* және *енгізу-шығару регистрі* терминдері өңделді, *сыртқы шығару* термині жеке термин ретінде емес, ол *саны* жалпы тұтынудағы сөз және *сыртқы шығару* сөздіктік емес терминдерін қосатын *сыртқы шығару саны* сөз тіркесінің бөлігі ретінде танылды - бұл қосылыс үшін стратегия ережелерінің ешқайсысы сәйкес келмеді.

Мұндай жағдайларды өңдеу стратегияға кірмейді, өйткені сөздік терминдерден тұратын сөз тіркестері көбінесе қосылыстар болып табылады (бірақ жеке терминдер емес) және Мcomb жиынтығының көптеген өңделмеген сөз тіркестерінде терминдер жоқ. Бұл жағдайды терминдерді алудың толықтығы аздап төмендейді.

4.7-кестеде қарастырылып отырған үзінді-мысал үшін терминдер мен оларды қолданудың толықтығы мен дәлдігі келтірілген:

* алу рәсімдері тәуелсіз жұмысының нәтижесі болып табылатын терминдер жиынтығын қарапайым біріктіру арқылы. Біріккен жиынтықтардағы терминдерді тұтыну жиілігін есептеу үшін және кірігуіне жол бермеу үшін getFrequency рәсімі қолданылады;
* стратегияға негізделген.

Кестеден стратегия терминдердің де, олардың әртүрлі тұтынылуының да толықтығы және дәлдігін айтарлықтай өсіретінін көруге болады.

Кесте 4.7 Мысал- мәтіннен алудың толықтығы мен дәлдігі

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Әдіс | Терминдерді алу | | | Тұтынуды алу | | |
| толықтығы | дәлдігі | F-өлшем | толықтығы | дәлдігі | F-өлшем |
| Бірлестік | 75,0% | 51,7% | 61,2% | 67,9% | 57,6% | 62,3% |
| Стратегия | 85,0% | 68,0% | 75,6% | 89,3% | 75,8% | 82,0% |

5-қосымшада бірнеше-теориялық операцияларды қолдана отырып, стратегия ережелерінің ресми жазбасы келтірілген: терминдік тұтыну рәсімдерінен алынған нәтижелер жиынтық ретінде қарастырылады, ал ережелер осы жиынтықтар мен олардың элементтеріне операциялар ретінде қарастырылады.

# 4.4 Қолданбалы тапсырмаларда алу рәсімдерін қолданылуы

Жеке алынған ҒТ-мәтінге терминологиялық талдау жүргізу талап етілетін МАӨ қолданбалы міндеттерінде әзірленген рәсімдерді қолдану жолдарын көрсетеміз. Осындай міндеттер ретінде ғылыми-техникалық құжаттың глоссарийі мен пәндік көрсеткішін құруды алайық.

Құжат глоссарийі - әрқайсысы анықтамамен қамтамасыз етілген құжаттың ең маңызды терминдерінің тізімі; терминдер үшін синонимдер, мәтіндік нұсқалар, мағыналық байланыстар да көрсетілуі мүмкін [78]. Глоссарийді құрастыру мәселесін толығымен автоматты түрде шешу өте қиын, өйткені глоссарийге енгізілуі керек құжаттың кейбір терминдері үшін мәтінде де, терминологиялық сөздіктерде де, тезаурустарда да нақты анықтамалар жоқ. Сондықтан, қазіргі уақытта осы қолданбалы мәселені шешуге арналған жұмыстардың басым көпшілігінде глоссарий құрудың автоматтандырылған режимі қабылданады [79]. Бұл жағдайда, әдетте, алдымен анықтамалары бар терминдер автоматты түрде шығарылады, содан кейін мәтіннің қалған терминдері (және оларды қолдану) анықталады, содан кейін сарапшы алынған ақпаратты өңдейді және толықтырады.

Біз жасаған рәсімдердің көмегімен құжаттың глоссарийін құру үшін келесі қадамдар тізбегі орындалуы керек:

1) авторлық терминдер мен олардың анықтамаларын шығару (getAuthTerms рәсімі).

2) авторлық терминдердің синонимдерін шығару (getSynTerms рәсімі).

3) авторлық терминдердің мәтіндік нұсқаларын және олардың синонимдерін шығару (getVarforTerms рәсімі).

Ұсынылған рәсімдер тізбегі [80] құжатқа қолданылды. Оның көмегімен табылған барлық үміткерлер глоссарий терминдері болып шықты. Табылмаған глоссарий терминдерінің мәтінде нақты анықтамалары жоқ. Салынған глоссарийге келесі элемент кірді, мысалы:

*CASE-құрал (CASE - ComputerAidedSoftwareEngineering) - ақпараттық жүйелерді құру және сүйемелдеу процестерін қолдайтын бағдарламалық құрал.*

Глоссарийге терминдердің мағыналық байланыстарын қосу қажет болған жағдайда, көрсетілген реттілік осы байланыстар тиісті қалыптар негізінде автоматты түрде алынатын қадаммен толықтырылуы мүмкін. Егер мағыналық қатынастарды алу қажет болса, мысалы, <бөлім-бүтін> сияқты типтік құралымдардың қалыптарын қолдану керек, атап айтқандаNl < CASE-құрал> «бірге» «бірге ««SPl<c=ins> «құрайды» SP2 < c=acc> # N1, SP2 cөйлемшедегі осы қалыптың көмегімен *CASE-құралдар жүйелік бағдарламалық жасақтамамен бірге АЖ әзірлеу ортасын құрайды, CASE -құрал (*бөлім*) және АЖ әзірлеу ортасы (*бүтін*) арасындағы байланыс алынады* [81]*.*

Сонымен қатар, глоссарий терминдерін қамтитын жаңа мағыналы сөз тіркестерін глоссарийге қосуға болады. Сонымен, *CASE-құрал* терминін қолданатын сөз тіркестерін шығару үшін, *іске асыру CASE-құралдар* және *пайдаланушы CASE-құралдар* мәтінінен терминдерді алуға мүмкіндік беретін N1 N2<CASE-құрал, c=gen> қалыбын қарастырған жөн.

Пәндік көрсеткішті құру міндеті анағұрлым күрделі болады. Пәндік көрсеткіш - бұл құжат терминдерінің және олар аталған беттердің құрылымдық тізімі [82]. Пәндік көрсеткішті құру үшін, біріншіден, құжаттың барлық терминдерін анықтау керек, екіншіден, мәтіндегі олардың әртүрлі кірістерін тану қажет. Автоматты түрде алынған ақпарат түзету үшін сарапшыға беріледі, өйткені бір жағынан пәндік көрсеткішке енгізілуі керек терминдердің көпшілігі мәтінде бір рет кездеседі және анықталмауы мүмкін (кейбір деректері бойынша мұндай терминдердің мөлшері 52% дейін [83]). Екінші жағынан, сілтемелер көрсеткіште белгілі бір термин қолданылатын барлық беттерге емес, тек талқылаудың негізгі тақырыбы болып табылатын беттерге орналастырылуы керек.

Пәндік көрсеткіштің құрылысын автоматтандыру үшін жоғарыда сипатталған терминдерді алу стратегиясын қолдануға болатынын эксперименттік зерттеу көрсетті. Стратегия олардан терминдер мен олардың әртүрлі қолданылуын алу үшін ғылыми-техникалық мәтіндер топтамасы қолданылды (4.2 бөлімін қараңыз). Стратегия алған нәтижелерді салыстырмалы бағалау үшін сол сынақ топтамасында терминдерді алудың ең танымал және жиі қолданылатын әдістері сыналды [84] :

* Pairs әдісі - терминологиялық екі сөзді тіркестерді (мүмкін қосымшамен) оларды құрайтын сөздердің бірге кездесу жиілігіне негіздей отырып алу;
* Terms әдісі – бір сөзді терминдер ескерілетін Pairs әдісінің модификациясы, сөздердің сөз таптарына шектеу енгізіліп, еркін ұзындықтағы сөз тіркестері ескерілетін;
* C-Value әдісі - терминологиялық тіркестерді олардың кездесу жиілігіне қарай алу;
* SP әдісі - терминдердің типтік синтаксистік үлгілерін қанағаттандыратын барлық сөз тіркестерін алу[85].

4.8-кестеде өңделген мәтіндік топтама үшін терминдер мен олардың тұтынылуын алудың толықтық, дәлдік және F-өлшемдік мәндері келтірілген.

Кесте4.8 Терминдерді алу стратегиясын тестілеу нәтижелері

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Терминдерді алу | | | Тұтынуды алу | | |
| Әдіс | толықтық | дәлдік | F-өлшем | толықтық | дәлдік | F-өлшем |
| Pairs | 27,3% | 13,0% | 17,6% | 24,4% | 20,4% | 22,2% |
| Terms-- | 54,1% | 37,4% | 44,2% | 69,2% | 41,5% | 51,9% |
| C-Value | 35,5% | 4,9% | 8,6% | 21,3% | 5,9% | 9,3% |
| SP | 51,4% | 22,6% | 31,4% | 37,3% | 29,7% | 33,1% |
| Стратегия | 53,6% | 73,1% | 61,8% | 68,1% | 59,7% | 63,6% |

Терминдерді алудың және оларды қолданудың ең жақсы толықтығын Terms әдісі көрсетті, оның көмегімен барлық бір сөздік терминдер мен ең жиі кездесетін көп сөздік терминдер танылды, бірақ оның қол жеткізген дәлдігі қанағаттанарлықсыз. Сонымен қатар, біздің стратегиямыздың көмегімен мәтінде қолдану жиілігіне қарамастан, бір сөзден аз, бірақ көп сөзден тұратын терминдер көбірек танылды. Осылайша, әзірленген стратегиямен көрсетілген толықтық Terms- әдісімен көрсетілген толықтықтан төмен емес, бірақ стратегиямен көрсетілген терминдерді алу дәлдігі және F-өлшемі және оларды қолдану шарасы айтарлықтай жоғары. Сонымен қатар, барлық дұрыс алынған терминдер үшін олардың әртүрлі қолданылуын танудың толықтығы мынаны құрайды: Terms әдісі үшін- 43,1% және біздің стратегиямыз үшін 48,4%. Осылайша, стратегия терминдердің барлық түрлерін толық анықтауға мүмкіндік береді.

# **ҚОРЫТЫНДЫ**

Диссертациялық жұмысты орындау барысында келесі нәтижелер алынды:

1. Терминологиялық сөз тіркестерінің құрылымын лексикалық-синтаксистік қалыптар түрінде біркелкі ресімдеуге мүмкіндік беретін тәсіл, сондай-ақ терминологиялық ақпаратты мәтіннен автоматты түрде алу рәсімдерінде қалыптарды пайдалану үшін оларды қолдану құралымдары мен нұсқалары ұсынылған.
2. Ұсынылған тәсіл аясында мәтіннен терминологиялық ақпаратты алу рәсімдері әзірленді, оның қалыптар түріндегі ресми сипаттамасына сүйене отырып және пайдаланылатын қалыптар жинағын өзгерту арқылы баптауға мүмкіндік береді.
3. Әзірленген рәсімдер бағдарламалық түрде жүзеге асырылады, терминдер қалыптарының, олардың нұсқалары мен оларды қолдану құралымдарының негізінде эксперименттік зерттеулер жүргізіледі.
4. Эксперименттік зерттеу нәтижелері бойынша мәтіннен терминдерді алудың дәлдігі мен толықтығы көрсеткіштерін жақсартуға мүмкіндік беретін іске асырылған рәсімдердің нәтижелерін біріктіру стратегиясы ұсынылды.

Жүргізілген эксперименттер қазақ тілді ғылыми-техникалық мәтіннен терминдер мен олардың қолданылуын алу F-өлшем шараларының жақсарғанын көрсетті.

Диссертацияда ұсынылған терминологиялық ақпаратты алу тәсілі, терминдерді автоматты түрде алу рәсімдері және рәсімдердің нәтижелерін біріктіретін стратегия терминдер мен оларды тұтынудың дәлдігі мен толықтығының көрсеткіштері маңызды болып табылатын МАӨ қолданбалы мәселелерін шешуде пайдалануға болады.

# **ПАЙДАЛАНЫЛҒАН ӘДЕБИЕТТЕР ТІЗІМІ**

1. Агеева Ю. А., Гудыма А. Л. Методы извлечения ключевых слов и терминов из текстов // Информационные технологии. – 2011. – № 6. – С. 36-43.
2. Колесникова О. Ю. Методы автоматического извлечения терминов из текста // Информационные технологии. – 2015. – № 8. – С. 29-35.
3. Брейман А. М., Кутуев Р. Ш., Новикова Ю. Н. Извлечение ключевых слов и терминов из текста с помощью машинного обучения // Вестник Южно-Уральского государственного университета. Серия: Вычислительная математика и информатика. – 2017. – Т. 6. – № 1. – С. 13-23.
4. Брыксин Т. А., Гуляев А. А., Потапенко А. А. Автоматическое извлечение терминов и их связей из текстов на естественном языке // Информационные технологии. – 2018. – № 6. – С. 426-431.
5. Шахова Е. И. Методы извлечения терминов из текстов: обзор и анализ // Известия высших учебных заведений. Поволжский регион. Технические науки. – 2019. – Т. 25. – № 2. – С. 45-56.
6. Зайцев М. В., Гаврилова Т. А., Камышева Е. В. Автоматическое извлечение терминов и их связей из текстов на русском языке // Информатика и ее применения. – 2019. – Т. 13. – № 4. – С. 54-65.
7. Васильева Е. В. Извлечение терминологии из текстов методами машинного обучения // Молодой ученый. – 2020. – № 4. – С. 225-229.
8. Сергиенко Л. С., Никитенко Е. В., Кирсанов Д. О. Методы извлечения терминов из текстов на русском языке // Научно-технический вестник информационных технологий, механики и оптики. – 2020. – Т. 20. – № 6. – С. 980-986.
9. Абдуллаева А. Х. Особенности терминов научно-технической лексики // Вестник Карагандинского университета. Серия: Филология. Педагогика. – 2011. – № 1. – С. 32-35.
10. Власенко Ю. А. Особенности функционирования терминов в научно-технических текстах // Вестник Московского университета. Серия 19: Лингвистика и межкультурная коммуникация. – 2014. – № 2. – С. 54-66.
11. Лукашевич Н. В. Особенности научно-технических терминов в русском и английском языках // Известия Самарского научного центра Российской академии наук. – 2015. – Т. 17. – № 6-5. – С. 1318-1321.
12. Рыбакова Ю. В. Особенности научно-технической терминологии в профессиональном общении // Язык и культура. – 2016. – № 1. – С. 8-15.
13. Мерзлякова Л. Н. Семантико-структурные особенности научно-технических терминов // Вестник Оренбургского государственного университета. – 2017. – № 9. – С. 95-100.
14. Полякова Е. В. Особенности использования научно-технических терминов в современном русском языке // Вестник Белгородского государственного технологического университета им. В. Г. Шухова. – 2018. – № 1. – С. 99-102.
15. Липунова М. И. Особенности научно-технической терминологии в контексте профессионального обучения // Научный диалог. – 2019. – № 7. – С. 280-291.
16. Коряковцева М. А. Особенности терминологии в профессиональных текстах научно-технической направленности // Вестник Пермского университета. Русский и иностранный язык. – 2020. – Т. 16. – № 3. – С. 335-348.
17. Арсентьев А. И., Клюшин Д. А. Формирование терминологического словаря на основе научно-технических текстов // Известия Томского политехнического университета. – 2007. – Т. 310. – № 5. – С. 168-172.
18. Бережной В. В., Булатова Т. В., Жаркова Е. В. Методика создания словарей терминологической информации на основе анализа научно-технических текстов // Наука и образование. – 2010. – № 9. – С. 123-128.
19. Гуляева Н. М. Методика формирования словаря терминов в области компьютерных наук // Сборник трудов Международной научно-технической конференции «Современные технологии в науке, образовании и производстве». – 2012. – С. 129-131.
20. Коваленко И. С. Методика формирования словаря терминов по экологической безопасности на основе анализа научно-технических текстов // Вестник Донского государственного технического университета. – 2015. – Т. 15. – № 2. – С. 231-236.
21. Носков А. А. Методика создания терминологических словарей в области гидроэнергетики // Вестник Челябинской государственной академии культуры и искусств. – 2017. – № 4. – С. 116-121.
22. Полякова Е. В. Методика формирования словаря терминов по научно-техническим специальностям // Научный диалог. – 2018. – № 6. – С. 266-273.
23. Черкашин В. Е., Каргаполова Е. В. Методика формирования терминологического словаря на основе анализа текстов научных статей // Вестник НГТУ. – 2019. – № 2. – С. 176-184.
24. Барахнин В. Б., Митренина О. А. Лексико-грамматические шаблоны в современном английском языке. – М.: Флинта, Наука, 2013. – 256 с.
25. Козлова И. В. Лексико-грамматические шаблоны в современном английском языке: семантика и функционирование. – М.: МГУ имени М. В. Ломоносова, 2015. – 220 с.
26. Фролова Н. Н., Решетникова О. В. Лексико-синтаксические шаблоны в современном английском языке // Известия Уральского федерального университета. Серия 2. Гуманитарные науки. – 2017. – Т. 17. – № 3. – С. 60-69.
27. Fries C. C. The Structure of English: An Introduction to the Construction of English Sentences. – Harcourt Brace Jovanovich, 1971. – 206 p.
28. Sinclair J. Corpus, Concordance, Collocation. – Oxford University Press, 1991. – 320 p.
29. Stubbs M. Text and Corpus Analysis: Computer-Assisted Studies of Language and Culture. – Blackwell, 1996. – 384 p.
30. Manning C. D., Schütze H. Foundations of Statistical Natural Language Processing. – The MIT Press, 1999. – 720 p.
31. Manning C. D., Raghavan P., Schütze H. An Introduction to Information Retrieval. – Cambridge University Press, 2008. – 496 p.
32. Biber D., Conrad S., Cortes V. If You Look at...: Lexical Bundles in University Teaching and Textbooks // Applied Linguistics. – 2004. – Vol. 25. – No. 3. – P. 371-405.
33. Wray A. Formulaic Language and the Lexicon. – Cambridge University Press, 2002. – 324 p.
34. Пархоменко, П. А. Обзор и экспериментальное сравнение методов кластеризации текстов / П. А. Пархоменко, А. А. Григорьев, Н. А. Астраханцев // Труды ИСП РАН. - 2017. - Т. 29, Вып. 2. - С. 161-200.
35. Агеев М. С., Добров Б. В., Лукашевич Н. В. Автоматическая рубрикация текстов: методы и проблемы // Учёные записки Казанского государственного университета. Серия Физико-математические науки. — 2008. — Т. 150, № 4. — С. 25–40.
36. Айсина Р. М. Обзор средств визуализации тематических моделей коллекций текстовых документов // Машинное обучение и анализ данных (http://jmlda.org). — 2015. — Т. 1, № 11. — С. 1584–1618.
37. Воронцов К. В. Аддитивная регуляризация тематических моделей коллекций текстовых документов // Доклады РАН. — 2014. — Т. 456, № 3. — С. 268–271.
38. Воронцов К. В., Потапенко А. А. Регуляризация, робастность и разреженность вероятностных тематических моделей // Компьютерные исследования и моделирование. — 2012. — Т. 4, № 4. — С. 693–706.
39. Дударенко М. А. Регуляризация многоязычных тематических моделей // Вычислительные методы и программирование. — 2015. — Т. 16. — С. 26–38.
40. Лукашевич Н. В. Тезаурусы в задачах информационного поиска. — Издательство МГУ имени М. В. Ломоносова, 2011.
41. Маннинг К. Д., Рагхаван П., Шютце Х. Введение в информационный поиск. — Вильямс, 2011.
42. Павлов А. С., Добров Б. В. Метод обнаружения массово порожденных неестественных текстов на основе анализа тематической структуры // Вычислительные методы и программирование: новые вычислительные технологии. — 2011. — Т. 12. — С. 58–72.
43. Крылова, И. В. Извлечение терминов из текстов научно-технической тематики / И. В. Крылова, О. А. Крупнова, С. А. Малых // Вестник Новосибирского государственного университета. Серия: Информатика, математика и механика. – 2012. – Т. 12, вып. 2. – С. 5-16.
44. Бритвин, О. А. Извлечение терминологии из текстов научно-технической тематики / О. А. Бритвин, А. Н. Виноградов // Информатика и ее применения. – 2012. – Т. 6, № 1. – С. 8-16.
45. Ковалева, Г. В. Методы и средства извлечения терминологии из текстов научно-технической тематики /Г. В. Ковалева // Информационные технологии и вычислительные системы. – 2014. – № 4. – С. 17-26.
46. Романенко, Е. И. Автоматическое извлечение терминов из текстов научно-технической тематики на основе морфологических признаков / Е. И. Романенко, А. И. Чернявский // Вестник компьютерных и информационных технологий. – 2015. – № 11. – С. 91-96.
47. Ступак, Е. И. Извлечение терминологии из текстов научно-технической тематики на основе ассоциативных связей / Е. И. Ступак // Системы и средства информатики. – 2016. – Т. 26, вып. 2. – С. 3-16.
48. Батырханов А.Г., Сагнаева С.К., Тусупов Д.А., Самбетбаева М.А. Тезаурустың классификациялық схемасының сипаттамасын формальдау тәсілі. // Хабаршы. Жаратылыстану және техникалық ғылымдар сериясы. Астана. ЕҰУ. 2017. №6(121). С.48-52.
49. Sambetbayeva M. A., Fedotova O. A., Fedotov A. M., Batyrkhanov A. G. Conceptual model of information system for resources storage. // Вестник. Восточно - Казахстанский государственный технический университет им. Д. Серикбаева. г.Уст-Каменогорск. 2018. №3. Том 1. Часть 3. С.13-26.
50. Самбетбаева М.А., Батырханов А.Г. Технология создания тематических интеллектуальных ресурсов для поддержки научно-образовательной деятельности на основе онтологического проектирования // IV международной научно-практической конференции "Информатика и прикладная математика", посвященной 70-летнему юбилею профессоров Биярова Т.Н., Вальдемара Вуйцика и 60-летию профессора Амиргалиева Е.Н. 25-29 сентябрь 2019, Алматы, Казахстан: Институт информационных и вычислительных технологий МОН РК. 2019. С. 270-280.
51. Самбетбаева М.А., Батырханов А.Г. Концептуальные подходы к построению распределенных систем для систематизации и классификации информационных ресурсов для научно-образовательной деятельности // Физика-математика ғылымдарының докторы, профессор Рамазанов Мұрат Ыбырайұлының 70 жылдық мерейтойына орайластырылған «Математика, механика жəне информатиканың теориялық қолданбалы мəселелері» Халықаралық ғылыми конференцияның материалдары 12–13 маусым: Академик Е.А. Бөкетов атындағы Қарағанды Мемлекеттік Университеті. Қарағанды. 2019. С. 146-147
52. Воробьев А.Е., Батырханов А.Г., Воробьев К.А. Основные подходы к поиску необходимой специализированной нефтяной информации // «Казахстанская нефть: прошлое, настоящее и будущее»: Сборник материалов Международной научно-практической конференции, посвященной 120-летию казахстанской нефти. - Атырау: Атырауский университет нефти и газа им. С. Утебаева. 2019. С. 321-326.
53. Воробьев А.Е., Батырханов А.Г., Воробьев К.А. Основы понятийной модели связи терминов нефтяного тезауруса // «Казахстанская нефть: прошлое, настоящее и будущее»: Сборник материалов Международной научно-практической конференции, посвященной 120-летию казахстанской нефти. - Атырау: Атырауский университет нефти и газа им. С. Утебаева. 2019. С. 326-331.
54. Kurmanbekovna, S. S., Batyrkhanov A., Aralbaevna, S. M., Bakirbaevna, S. Z., Sembekovna, Y. A. Development of technology to support large information storage and organization of reduced user access to this information. International Journal of Advanced Computer Science and Applications, 2021, 12 (7), 493–503. doi: https://doi.org/10.14569/ijacsa.2021.0120757
55. Serikbayeva, S., Tussupov, J., Sambetbayeva, M., Yerimbetova, A., Sadirmekova, Z., Tungatarova, A., Batyrkhanov, A., Zakirova, A. Development of model and technology access to documents in scientific and educational activities. Eastern-European Journal of Enterprise Tech-nologies, 2021, 6 (2 (114), Р.44-58. https://doi.org/10.15587/1729-4061.2021.248506
56. Batyrkhanov, A.G., Sadirmekova, Z.B., Sambetbayeva, M.A., Nurgulzhanova, A.N., Ismagulova, Z.S., Yerimbetova, A.S. Development of methods and technologies for creating intelligent scientific and educational internet resources. Bulletin of Electrical Engineering and Informatics, 2022, 11 (5), pp. 2968-2977
57. Леонова Ю.В. Технология построения информационно-справочных систем для научного сообщества: автореф. … канд. техн. наук: 05.25.05. – Новосибирск, 2005. – 17 с.
58. Антопольский А.Б. Алгоритмы и процедуры определения уровня развития информационных ресурсов в научных и образовательных организациях // Науч.-техн. информация. – 2015. – №1. – С. 14-24.
59. Антопольский А.Б. Лингвистическое обеспечение электронных библиотек: учеб. пос. – М., 2003. – 301 c.
60. Антопольский А.Б. Зарубежные и российские ресурсы в сфере научного наследия и основные принципы создания электронной библиотеки «Научное наследие России» // Профессиональные штудии: информ. сб. – М., 2012. – Вып. 3 – С. 20-21.
61. Баженов С.Р., Жижимов О.Л., Поздняков И.Л. Актуальные проблемы доступа к БД посредством протокола Z39.50 // Библиотеки и ассоциации в меняющемся мире: новые формы технологии и новые формы сотрудничества: матер. конф. – М., 2001. – С. 775-780.
62. Малицкий Н.А., Елепов Б.С. Функциональные возможности современных библиографических ИПС на Web// Научные и технические библиотеки. – 2005. – №2. – С. 50-55.
63. Мазов Н.A. Методическое пособие по курсу Система CDS/ISIS в корпоративных библиотечных системах. –Новосибирск, 2001. – 140 с.
64. Khamdamov U., Abdullayev A., Elov J. et al.Conceptual Model of the Education Management Information System for Higher Education Institutions // International Journal of Advanced Trends in Computer Science and Engineering.– 2020. – Vol.9(5). – P. 7295-7300.
65. Molnár B., Benczúr A.Modeling information systems from the viewpoint of active documents // Vietnam Journal of Computer Science. – 2015. – Vol. 2(4). – P. 229-241.
66. Подкорытова Н.И., Лакизо И.Г., Артемьева Е.Б. Научные библиотеки в научно-образовательном комплексе Новосибирска // Научные и технические библиотеки. – 2020. – №1(9). – С. 61-77.
67. Белоногов Г.Г., Кузнецов Б.А. Языковые средства автоматизированных информационных систем. -М.: Наука, 1983.
68. Большакова Е.И., Баева Н.В., Бордаченкова Е.А., Васильева Н.Э., Морозов С.С. Лексико-синтаксические шаблоны в задачах автоматической обработки текстов // Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции Диалог'2007. - М.: Изд-во РГГУ, 2007. - Т. 2. - С. 70-75.
69. Большакова Е.И., Баева Н.В., Васильева Н.Э. Структурирование и извлечение знаний, представленных в научных текстах // Девятая Национальная конференция по искусственному интеллекту с международным участием КИИ-2004. Труды конференции. - М.: Физматлит, 2004. - Т. 2. - С. 480-488.
70. Большакова Е.И., Васильева Н.Э. К вопросу об автоматизации литературно-научного редактирования // Компьютерная лингвистика и ее приложения: Труды Международного семинара Диалог'2000. - Протвино, 2000. - Т.2. - С. 59-63.
71. Большакова Е.И., Васильева Н.Э. Терминологическая вариантность и ее учет при автоматической обработке текстов // Одиннадцатая Национальная конференция по искусственному интеллекту с международным участием КИИ- 2008. Труды конференции. -М.: Физматлит, 2008. - Т. 2. - С. 174-182.
72. Большакова Е.И., Васильева Н.Э. Формализация лексико- синтаксической информации для распознавания регулярных конструкций естественного языка // Программные продукты и системы. - 2008. - № 4. - С. 103-106.
73. Большакова Е.И., Васильева Н.Э., Морозов С.С. Лексико- синтаксические шаблоны для автоматического анализа научно-технических текстов // Десятая Национальная конференция по искусственному интеллекту с международным участием КИИ-2006. Труды конференции. - М.: Физматлит, 2006.-Т. 2.-С. 506-524.
74. Большакова Е.И., Васильева Н.Э., Юдин Д.А. Выделение словарных терминологических словосочетаний в научно-технических текстах // Компьютерная лингвистика и ее приложения: Труды Международного семинара Диалог'2001. - Аксаково, 2001. - с. 48-51.
75. Большакова Е.И., Ефремова Н.Э., Носков A.A. Методы и средства построения программных систем для анализа текста с использованием лингвистических шаблонов // Ломоносовские чтения: научная конференция, посвященная 300-летию со дня рождения М.В. Ломоносова: Тезисы докладов. - 2011.-С. 97.
76. Борковский А.Б., Зайчик Б.И., Боровикова Л.И. Словарь по программированию. - М.: Русский язык, 1991.
77. Браславский П.И., Соколов Е.А. Сравнение четырех методов автоматического извлечения двухсловных терминов из текста // Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции Диалог'2006. - М.: Изд-во РГГУ, 2006. - С. 88-94.
78. Браславский П.И., Соколов Е.А. Сравнение четырех методов автоматического извлечения двухсловных терминов из текста // Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции Диалог'2006. - М.: Изд-во РГГУ, 2006. - С. 88-94.
79. Добров Б.В., Лукашевич Н.В., Сыромятников C.B. Формирование базы терминологических словосочетаний по текстам предметной области // Труды пятой всероссийской научной конференции "Электронные библиотеки: Перспективные методы и технологии, электронные коллекции". - 2003.-С. 201-210.
80. Емашова O.A., Мальковский М.Г. Функциональные стили русского языка и их влияние на задачу автоматического реферирования текста // Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции Диалог'2007. - М.: Изд-во РГГУ, 2007. - Т.2. - С. 167-171.
81. Ефремова Н.Э., Большакова Е.И., Носков A.A., Антонов В.Ю. Терминологический анализ текста на основе лексико-синтаксических шаблонов // Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции Диалог'2010. - М.: Изд-во РГГУ, 2010. - С. 124­129.
82. Жигалов В.А., Жигалов Д.В., Жуков A.A. и др. Система Alex как средство для многоцелевой автоматизированной обработки текстов // Компьютерная лингвистика и интеллектуальные технологии: ТрудыМеждународной конференции Диалог'2002. - М.: «Наука», 2002. - Т. 2. - С. 192-208.
83. Леонтьева H.H. Автоматическое понимание текстов: системы, модели, ресурсы: учебное пособие для студентов лингвистических факультетов вузов. - М.: Издательский центр "Академия", 2006.
84. Носков A.A. Метод выделения в тексте конструкций по их лексико- синтаксическим шаблонам // Сборник статей молодых ученых факультета ВМиК МГУ. - М.: Издательский отдел фак-та ВМиК МГУ имени М.В. Ломоносова; МАКС Пресс, 2009. - Вып. 6. - С.136-145.
85. Скатов Д.С., Вдовина Н.А., Ливерко C.B. и др. Язык описания правил в системе лексического анализа ЕЯ-текстов DictaScope Tokenizer // Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции Диалог'2010. - М.: Изд-во РГГУ, 2010. - С. 442­449.

ҚОСЫМША А

Лексикалық-синтаксистік қалыптар тілінің синтаксисі

Төмендегі БНТ формулаларында қалып тілінің бейтерминалдары курсивпен, терминалдар қалың әріппен жазылған.

қалып ::= [қалып атауы =]қалып денесі{| қалып денесі}

[(қалып параметрі)][# қалыптар^алу]

қалып атауы ::= Бас. лат. әрпі {лат. әрпі}

қалып денесі ::= қалып элементі {қалып элементі)

[<шарттар>]

қалып элементі\_ ::= қарапайым\_элемент|опционалды\_элемент|

міндетті \_элемент

қарапайым\_элемент : := жол-элементі | сөз-элементі |

қалып данасы

опционалды элемент ::=

[балама\_элемент]міндетті\_элемент ::=

{ балама \_элемент}балама\_элемент ::= қарапайым\_элемент {| қарапайым\_элемент} жол-элементі ::= "символ {символ}"

сөз-элементі ::= сөз-элемента атауы | сөз-элементі\_атауы

<сөз\_үлгісі {, сөз\_сипаттамасы} > |сөз- элементі\_атауы < сөз сипаттамасы

{, сөз\_сипаттамасы) >

сөз-элементі\_атауы ::= сөз\_табы [индекс]

сөз\_табы ::=W | N | А | V | Ра | Ар | Рn | Av | Сn | Рг | Pt | In |Nm

индекс ::= сан {сан}

сөз үлгісі ::= лексема\_атауы

лексема\_атауы::=қазақ әрпі {қазақ әрпі}

қазақ әрпі::= кириллица\_әрпі | дефис

сөз\_сипаттамасы::= белгі\_ атауы = белгі\_ мәні

белгі\_ атауы::= с | n | g | doc | a|f|m|p|r

белгі\_ мәні::=nom | gen| dat| асс | ins| prep| un | sing|

plur|masc| fem | neut | comp | sup | près |

past |fut | inf |nim | inan | full | short |

ind | imp | conj |cond | 1 | 2 | 3 | yes | no

қалып \_данасы ::= қалып\_данасы\_атауы

[<дана сипаттамасы

{, дана\_сипаттамасы}>]

қалып\_данасы\_атауы::= қалып\_атауы[индекс] дана\_сипаттамасы ::= белгі\_атауы = белгі\_мәні

шарттар ::= келісімшарттары | сөздіктік\_шарттары

келісім\_ шарттары ::= келісім {, келісім}

келісім ::= = элемент\_атауы |

элемент\_атауы. элемент\_атауы = элемент атауы. белгі\_атауы

элемент\_атауы::= сөз-элементі\_атауы | қалып\_данасы\_атауысөздіктік\_шарттары::= сөздіктік\_шарттары

{, сөздіктік^шарттары }

сөздіктік\_шарттары::= сөздік\_атауы (элемент\_атауы {, элемент\_атауы }) қалып\_параметрі ::= элемент\_атауы {, элемент\_атауы }

алу­­\_қалыптары::= қалып\_денесі {| қалып\_денесі }

А. 1-кестесінде қазақ тіліндегі сөздердің сөз таптары белгілері және олардың морфологиялық сипаттамалары келтірілген.

Кесте-А. 1. Сөз таптары және олардың морфологиялық сипаттамалары

|  |  |  |
| --- | --- | --- |
| Сөз таптары (part) | Белгісі | Мүмкін сипаттамалар |
| Еркін сөз табы (word) | W |  |
| Зат есім (noun) | N | жағы, жанды-жансыздығы, септігі, саны |
| Сын есім (adjective) | А | толық, қысқа, шырай түрі немесе жай шырай, септігі (толығында), сан (толығы мен қысқасында), жағы (жекеше түрдегі толығы мен қысқасында) |
| Етістік (verb) | V | рай, шақ (ашық райда), жақ, саны (ашық райда, өткен шақта, жекеше түрде), болымды |
| Есімше (participle) | Ра | Толық және қысқа, септігі (толығында), саны, жағы (жекеше түрде), болымды |
| Көсемше (adverbial participle) | Ар | шақ, болымды |
| Есімдік (pronoun) | Рn | септік, саны, жағы (мүмкін болса) |
| Сан есім (numeral) | Num | септігі, жағы |
| Үстеу (adverb) | Av­ |  |
| Шылау(conjunction) | on |  |
| Қосымша (preposition) | Pr |  |
| Бөлшек (particle) | Pt |  |
| Одағай (interjection) | Int |  |

А.2-кестесінде морфологиялық сипаттамалардың мүмкін мәндері және олардың қалып тіліндегі белгілері келтірілген

Кесте- А.2. Морфологиялық сипаттамалары және олардың мағыналары

|  |  |  |
| --- | --- | --- |
| Белгі және оның атауы | Қысқарту | Сипаттамалардың мүмкін мәндері және олардың белгілері |
| Септік (case) | с | Атау (nominative) - nom |
|  |  | Ілік (genitive) -gen |
| Барыс (dative) -dat |
| Табыс (accusative) - **асс** |
| Көмектес (instrumental) -ins |
| Жатыс (prepositional) -prep |
| Өзгермейтін (uninflected) –un |
| Саны (number) | п | Жекеше түр (singular) –sing |
|  |  | Көпше түр (plural) –plur |
| Салыстырма шырайы (degree of comparison) | doc | Салыстырмалы (comparative) –com |
| Күшейтпелі (superlative) –sup |
| Жай (no) –no |
| Шақ (tense) | t | Осы (present) –pres |
|  |  | Өткен (past) –past |
| Келер (future) –fut |
| Тұйық етістік (шақсыз) (infinitive) -inf |
| Жанды-жансыздығы (animate) | а | Жанды (animate) –anim |
| Жансыз (inanimate) -inan |
| Тұлғасы (form) | f | Толық (full) –full |
|  |  | Қысқарған (short) –short |
|  |  | Қысқарған (short) –short |
| Райы (mode) | m | Ашық (indicative) –ind |
|  |  | Бұйрық (imperative) –imp |
| Қалау (conjunctive) -conj |
| Шартты (conditional) -cond |
| Жақ (person) | Р | 1 жақ -1 |
|  |  | 2 жақ - 2 |
| 3 жақ - 3 |
| Болымды-болымсыздығы (reflexive) | г | Болымсыз (no) -no |
| Болымды (yes) -yes |

# **ҚОСЫМША Ә**

Терминдерді анықтау қалыптары және олардың синонимдерін ендіру

Б. 1-кестесінде ғылыми-техникалық терминдердің анықтамаларының ең типтік құралымдарының үлгілері келтірілген. Қалыптар олардағы лексемалар бойынша топтастырылған. Қалыптар мен мысал-үзінділерде авторлық терминдердің асты сызылады.

Кесте- Б.1. Терминдердің анықтама қалыптарының мысалдары

|  |  |
| --- | --- |
| Терминдердің анықтама қалыптары | Терминдердің анықтама мысалдары |
| V< аталу> | |
| Term<біргі=ins> V<аталу>, t=pres,р=3,m=ind>Defin<с=nom><Term.n=V.n># Term | Циклдің периодтық кестесі деп Г... аталады. |
| Defin<с=nom> [Prep<с=prep>] V<аталу>,t=pres,p=3,m=ind> Term<с=ins><Defin.n=V.n># Term | Бұл әсер геодезиялық сызықтар девациясы деп аталады |
| TermN<с=nom> V<аталу>, t=pres,p=3,m=ind>TermA<с-ins><TermN.g=TermA.g, TermN.n=TermA.n=V.n>  # TermATermN<TermA=TermN> | Мұндай мюондар альбедті деп аталады.  Келтірілген қоздыру әдісі контурлық де аталады. |
| Ра< атау> | |
| Defin{"," |"(" } ["әдетте"]  Ра<атау> ["жиі"]  Term<с=ins><Pa=Defin># Term | Жұмыс әдетте сегменттеу әдісі деп атайтын тәсілдерді зерттеуге арналады |
| "осылай" Ра<атау>Term  <Pa=Term># Term | Жеделдетудің серфотрондық механизмі деп атайтын таласады |
| "атайық" | |
| "атайық" Defin<с=acc>  Term<с=ins># Term | Осы түрленуді градиентті интегралды түрлену деп атайық |
| Ра< атау> | |
| Defin{","|"("} Ра<атау>  Term<с=ins><Defin=Pa># Term | Аралдық кластерлеу деп аталған әдістерді бірі деп қабылдайық |
| Терминдердің анықтама қалыптары | | Терминдердің анықтама мысалдары |
| V<алу> " атауы " | | |
| Defin<с=nom>  V<алу,t=past,р=3,m=ind>  "атауы" Term<с=gen>  <Defin.n=V.n,Defin.g=V.g># Term | | Сипатталға қасиеті бар элементтер полимерлік демпферлер деген атау алды. |
| V<ұғыну> | | |
| "деп" Term<с=ins>[Prep<с=prep>] V<ұғыну,t=pres,p=3,m=ind>Defin<с=nom><Term.n=V.n># Term | | Дұрыс қолдану деп, осы жағдайда функцияның қолданылуы ұғынылады |
| "деп" "боламыз" "түсіну" | | |
| "дп" Term<с=ins>  ["бі"] "боламыз" "түсіну"  Defin<с=acc># Term | | Деректердің интелектуалдық қоры деп ...қорын түсінетін боламыз |
| "деген" | | |
| Term<с=nom>["-"] "деген"  Defin<с=nom># Term | | Сандық жазықтық дегеніміз бар сандардың барлық жиынтығы |
| "я" "яғни", "яки" | | |
| Term{" (" | "," }  {"я" "деген" | "яки."} Defin<Term.с=Defin.с,Term.n=Defin.n>  # Term | | Жол дөңгелектеулеріндегі қисықтар шеңбер доғасы, яғни тұрақты қисықтың қисықтығымен сипатталған |
| "бұл" | | |
| Term<с=nom>["-"] "бұл"  Defin<с=nom># Term | | Жеткіліксіз масса – бұл суыған қарапайым нысан массасы |

Б. 2-кестеде ғылыми-техникалық терминдердің синонимдерін енгізудің ең типтік құралымдарының қалыптары келтірілген. Қалыптар мен мысал-үзінділерде терминдер мен олардың синонимдерінің асты сызылған.

Кесте- Б.2. Синонимдерді енгізу қалыптарының мысалдары

|  |  |
| --- | --- |
| Синонимдерді енгізу қалыптары | Синонимдерді енгізу мысалдары |
| SP1 "(" ["немесе"]SР2 ")"  <SР1.с=SР2.с> # SР1, SР2 | ...баспаланған арнайы салыстырмалылық теориясын (АСТ) жетілдіретін геометриялық теория |
| Синонимдерді енгізу қалыптары | | Синонимдерді енгізу мысалдары |
| SP1 "," "немесе" ["қарапайым"] SP2 # SP1, SP2 | | ..дәстүрлі түрде интервал немесе әлемдік функция деп аталатын |
| AP1 "(" ["немесе"]AP2 ")" PSP1 <AP1=AP2=PSP1>#  AP1 PSP1 <AP1=PSP1>,  AP2 PSP1 <AP2=PSPl> | | ...тензорлық (инвариантты) нәтижелер алу ... математикалық тұрғыдан мүмкін емес |
| " (" "ары қарай" ["-"]  Term<с=nom>")" # Term | | ......бастапқы саланың бұдан әрі (БС) барлық тұжырымдамалары (бұдан әрі олар) мақсат-саланың (МС) тұжырымдамаларымен сәйкес келеді... |

Терминдердің анықтамалары мен олардың синонимдерін енгізу қалыптарында келесі көмекші қалыптардың даналары қолданылады:

Term= SP1 [[","] "немесе" ["қарапайым"] SP2]

<SPl.с=SP2.с> (SP1) # SP1, SP2

TermA= API [[","] "немесе" АР2 [","]]

<APl=AP2> (API) # API, АР2

TermN= PSP [PG] <PSP=PG> (PSP) # PSP

PSP = N1 [N2<с=gen>] (N1)

SP = N1 (N1) | N1 N2<с=gen>(N1) | API N1 <APl=Nl>(N1) |

AP1 AP2 N1 <AP1=AP2=N1>(N1) |

AP1N1 N2<с=gen><AP1=N1>(N1) |

N1 AP1 N2<с=gen><AP1=N2>(N1) |

N1 N2<с=gen> N3<с=gen>(N1)

Defin= {Рп<кез келген>|Рп<осындай>|Рп<әрбір>|Рп<бұл>}

SP <Pn=SP> (SP) | SP [PG <SP=PG>|SG <SP=SG>] (SP) |

"онда" |"бұл"

PG - "," ["жоқ"] Pa {WD} "," (Pa)

SG = "," [Pr] Рп<қайсы>{WD} "," (Pn)

WD = W | Del

Del = "," | "-"

AP = A<f=full> (A) | Pa<f=full> (Pa)

Prep= PrSP (SP)

# **ҚОСЫМША Б**

Сөздіктік қалыптар мысалдары

**Информатика мен есептеу техникасы (ИмЕТ) бойынша терминологиялық сөздік үлгілері**

абонент = N1<абонент>N2<желілер, с=gen>

апат = N1<апат>

база = N1<база>

[N2<деректер, с=gen,п=plur>

[А3<ұжымдық>NЗ<қолдану, с=gen><А3=N3>| А3<жалпы> {N3<тағайындау, с=gen>

N3 пайдалану,с=gen>} <АЗ=NЗ>|

А3<пәндік>N3<сала,с=gen><АЗ=NЗ>| N3<кәсіпорын,с=gen>]| N2<білім,с=gen,n=plur>]

ығысу-база = N1< ығысу-база>

**Физика бойынша терминологиялық сөздік қалыптары**

n-нұсқалық = А1<n-нұсқалық> А2<термодинамикалық>

N1<жүйе><А1=А2=:N1>

жүгіру = Ра1<жүгіру>N1<толқын><Ра1=N1>

кеңістіктік = А1<кеңістіктік>

{N1<дисперсия>|N1<инверсия>|N1<когеренттік>

N1<модуляция>|N1<кванттау>|N1<қуат>} <А1=N1>

релаксация = N1<релаксация> [N2<кернеу, с=gen,n=plur>]

**ИмЕТ саласының термин-дублет жұптарының қалыптары**

N1<бит> # N2<разряд>

N1<конъюнкция> # [N2<операция>] "және"

А1<арифметикалық-логикалық>N1<құрылғы><А1=N1>

# А2<арифметикалық>N2<құрылғы><А2=N2>, "АЛҚ", "АҚ"

А1<кассетті>N1<жинақтауыш> Рг1<оған>N2<таспа><А1=N1>

# А3<кассетті>N3<жинақтауыш> Рг2<оған> А4<магнитті>

N4<таспа><АЗ=N3,А4=N4>

А1<құрамдық>N1<мағынасы><А1=N1> # N2<агрегат>

**Физика саласындағы термин-дублет жұптарының қалыптары**

А1<электрлік>N1<вектор><А1=N1>

#А2<жарықтық>N2<вектор><А2=N2>

А1<электрлік>N1<сыйымдылық><А1=N1>

# N2<электрсыйымдылық>

N1<кедергі> "Шотки" # N2<кедергі> "Шотки"

N1<жүйе>N2<бірлік,с=gen,n=plur> "СГСФ" # "СГСФ"

А1<цилиндрлік> А2<магниттік>N1<домен, n=plur>

<А1=А2=N1> # "ЦМД", А3<магниттік>N3<көпіршік><А3=N3>

**ИмЕТ саласының синонимдер сөздігі қалыптары**

N1<деректер, п=р1иг> # N2<ақпарат>

N1<жинау> # N2<құрастыру>

А1<логикалық> # А2<бульдік>

N1<тапсыру>N2<басқару, с=gen> # N3<ауысу>

N1<конъюнкция> # А2<логикалық>N2<көбейту><А2=N2>

**Физика саласындағы синонимдік сөздік қалыптары**

А1<гомеполярлы> # А2<гомполярлы>, А2<ковалентті>

N1<пи-мезон,n=plur> # N2<пион, n=plur>

N1<импульс> # N2<мөлшер>N3<қимыл, с=gen>

А1<сәулелі>N1<ағым><А1=N1> # n2<сәулелену>

**ИмЕТ саласынан алынған дериваты сөздік қалыптары**

А < мекенжайлық> # N < мекенжай>

A < интерфейс> # N < интерфейс>

А < бағдарламалық> # N < бағдарлама>

**Физика саласындағы дериват сөздік қалыптары**

A < векторлық># N<вектор>

A < құйындық> # N < Құйын>

A < толқындық> # N < толқын>

# **ҚОСЫМША В**

Терминдердің қолданылу нұсқаларының қалыптары

Лексикалық-синтаксистік нұсқаларды құру ережелерінде қалыптардың денелері (терминологиялық сөз тіркестерінің типтік синтаксистік үлгілеріне сәйкес) қалың әріппен ерекшеленеді. AP1N1 және N1 N2<c=gen > үлгілері үшін олар сипаттайтын терминдер мен нұсқалардың мысалдары келтірілген. Шылаулы және шылаусыз қосылыстардан терминдерді алу ережелерінде барлық ережелер үшін мысалдар келтірілген.

**Лексикалық-синтаксистік нұсқаларды қалыптастыру ережелері**

AP1N1 <AP1=N1># N1

Формалды сипаттама - сипаттама

AP1N1 <AP1=N1># АР2 N1 <Syn(АР2,API),AP2=Nl>

ақпараттық технологиялар - компьютерлік технологиялар

AP1N1 <AP1=N1># AP1N2 <Syn(N2,N1),APl=N2>

құрылымдық талдау – құрылымдық әдістеме

AP1N1 <AP1=N1># N1 N2<c=gen><Der(API,N2)>

Блоктық жіберу – блокты жіберу

N1 N2<с=gen># N1

пайдаланушы CASE-құрал – пайдаланушы

N1 N2<с=gen># N1 N4<с=gen><Syn(N2,N4)>

қуат тасушы — токты тасушы

N1 N2<с=gen># N3 N2<с=gen><Syn(Nl,N3)>

талап тұжырымдамасы — талап сипаттамасы

N1 N2<с=gen># APIN1 <Der(API,N2),AP1-N1>

Бор радиусы — борлық радиус

AP1АР2 N1 <AP1=AP2=N1>

# N1, АР2 N1 <AP2=N1>, AP1N1 <AP1=N1>,

АР3AP2 N1 <Syn(AP3,API),AP2=AP3=N1>,

AP1 AP3 N1 <Syn(АРЗ,AP2),AP1=AP3=N1>,

AP1 AP2 N2 <Syn(N1,N2),AP1=AP2=N2>,

AP1N2 <Syn (AP2 N1,N2),APl=N2>

API N1 N2<с=gen><AP1=N1>

# N1, API N1 <AP1=N1>, N1 N2<с=gen>,

АР3 N1 N2<с=gen><Syn(АРЗ,Al),AP3=N1>,

AP1N3 N2<с=gen><Syn(N3,N1),AP1=N3>,

AP1N1 N3<с=gen><Syn(N3,N2),AP1=N1>,

AP1N3 <Syn(N3,N1 N2),APl=N3>,

N3 N2<с=gen><Syn(N3,APl Nl)>

N1 N2<с=gen> N3<с=gen>

# N1, N1 N2<с=gen>,

N4 N2<с=gen> N3<с=gen><Syn(N4,N1)>,

N1 N4<с=gen> N3<с=gen><Syn(N4,N2)>,

N1 N2<с=gen> N4<с=gen><Syn(N4,N3)>,

N4 N3<c=gen><Syn(N4,Nl N2)>,

N1 N4<c=gen><Syn(N4,N2 N3)>

N1 AP2 N2<c=gen><AP2=N2>

* N1, AP2 N1<AP2=N1>,

N3 AP2 N2<c=gen><Syn(N3,N1),AP2=N2>,

N1 AP3 N2<c=gen><Syn(АРЗ,AP2),AP3=N2>,

N1 AP2 N3<c=gen><Syn(N3,N2),AP2=N3>,

N1 N3<c=gen><Syn(N3,AP2 N2)>

**Шылаулы қосылыстардан терминдерді алу ережелері**

[AP1","] АР2 {"және"|"немесе"|"әлде"} АРЗ N1

<АР1.с=АР2.с=АРЗ.c=Nl.с, API.g=AP2.g=AP3.g=Nl.g,

AP1.n=AP2.n=AP3.n>

* AP1N1 <AP1=N1>, AP2 N1 <AP2=N1>, AP3 N1 <AP3=N1>

...тағайындалған ... жұмыс және графикалық станциялар үшін... – жұмыс станциясы, графикалық станция

[N1 ","] N2 {"және"|"немесе" |"әлде"} N3 N4<c=gen>

<N1.n=N2.n=N3.n, N2.c=Nl.c=N3.c>

* N1 N4<c=gen>, N2 N4<c=gen>, N3 N4<c=gen>

...құралдар, қолдайтын ... талаптарды талдау және тұжырымдау... – талаптарды талдау, талаптарды тұжырымдау

N1 N2<c=gen>["," N3<c=gen>] {"және" Г'немесе"|"әлде"} N4<c=gen>

* N1 N2<c=gen>, N1 N3<c=gen>, N1 N4<c=gen>

... микропроцессор мекенжай, деректер және басқару құрсымдарының - мекенжай құрсымы, деректер құрсымы, басқару құрсымы арқасында құрылғылармен "сөйлесе" алады...

**Шылаусыз қосылыстардан терминдерді алу ережелері**

AP1 АР2 N1 <AP1=AP2=N1># API N1 AP1=N1>, AP2 N1 <AP2=N1>

...ұзын бүтін сандарды көрсету үшін қосымша екілік таңдалады... - ұзын сан, бүтін сан.

N1 АР2 N2<c=gen><AP2=N2># АР2 N2 <AP2=N2>, N1 N2<c=gen>...сыртқы талаптарды сипаттауға арналған диаграммалар немесе мәтіндер түріндегі сипаттамалар... - талаптардың сипаттамасы, сыртқы талаптар.

# **ҚОСЫМША Г**

Стратегия ережелері

Терминдерді алу стратегиясының жұмыс ережелерін сипаттау кезінде біз келесі белгілерді қолданамыз. termi арқылы getAuthTerms, getDictTermsжәне getNonDictTerms(мысалы, Mauth= {term1, term2, ...,terma}) рәсімдері арқылы қайтарылатын үміткер-терминдер жиынтығының элементтерін белгілейік. Кез келген termi бүтін сандардың көптеген жұптарын салыстырады-барлық termi талдайтын мәтінге кіру шекараларын, біз оны келесідей белгілейміз termi={ (begil, endi1) , (begip, endip) , ..., (begip, endip) }.

Ережелерді орындауды бастамас бұрын М:=0.

1-ереже.Айталық M:=MMauth.

2-ереже. Әрбір termdMdict үшін тексереміз: егер үшін termnMnondict termdtermn болса, онда M:=M {termd} болады.

Mcomb= {comb1, comb2, ..., combc} шылаусыз қосылыстар жиынтығы екі жиынтыққа бөлінген:

* Мсогг үзілген не/немесе қысқарған терминдердің қосылыстары;
* Msave.үзілмеген не/немесе қысқармаған терминдердің қосылыстары

Әр қосылыс үшін оның құрамына кіретін үміткер-терминдер анықталған: combi= {termi1, termi2, ..., termip}

3-ереже. Әрбір elemnMnondict үшін тексереміз:

* 1. Егер combnMcorr және termdMdict (мүмкін біреу емес) осындай болса, termdcombn болғанда, онда M:=M{termd} болады.
  2. Егер combnMsave\Mcorr және S= {termnj|termnjMdictcombn} осылай болса, |S | >1 болғанда, онда M:=MS болады.

(3) Егер combnMsave\Mcorr және 3 ! termEcombn осылайша, ал termEMdlct және termCcombn болғанда, сондай-ақ басқалардың бәріне termn:iEcombn дұрыс болса, termn-,£MUMnondlct болғанда, онда M:=MU {combn} болып, ал Mnondlct:=Mnondlct\{combn} және Msave:=Msave\ { combn} болады.

Ережені қолданғанда соңында Mdlct:=Mdlct\M болады.

Msyn={ (syn11, syn12) , (syn21,syn22),..., (synn1, synn2) } – синоним жұптарының жиынтығы; Msyn­\_br={syn1;syn2, ...,syns}- енгізу құралымдарында үзілген түрде кездескен синоним жиынтығы. Айталық С : = | М |.

4-ереже. Әрбір (syni1, syni2) жұбы үшін тексереміз:

* + 1. Егер syni1 M, ал synl2M, онда M:=M {synl2}, және керісінше, егер synl2M, asynilM, онда M:=M{syni1};Msyn:=Mgyn\{ (syni1, synl2) }.
    2. Егер syni1 M және synl2M, бірақ syni1 Mdict Msyn\_br немесе synl2Mdict Msyn\_br, онда M:=M{syni1, syni2}; Msyn:=Msyn\{ (synn, synl2) }.

Mconj={comb1, comb2, ..., сombt}- шылаулы қосылыстар жиынтығы, әр қосылыс үшін оның құрамындағы үміткер-терминдер анықталған: comb1={termi, termi2, termip}.

Mconj­\_br={term1, term2, ..., termb} – шылаулы қосылыстарда үзілген түрде кездескен үміткерлер жиынтығы.

5-ереже.Әрбір combcMconj үшін тексереміз: егер кем дегенде біреуі termccombc осылайша болса, termcj M Mdlct немесе termcj Mnondict Mconj\_brr болғандықтан, онда M:=MU{termc1, termc2, ..., termcp}, ал Mdlct:=Mdlct\{termcl, termc2, ..., termcp} және Mnondict:=Mnondict\{termc1, termc2, ..., termcp}; Mconj:=Mconj\ {combc} болады.

6-ереже.Әрбір cornbcMcomb үшін тексереміз:

* + - 1. Егер combcMcorr және termcj combcMcomb\_br осылай болса, termcj MUMdlctUMnondlct болғандықтан, онда M:=MU {termcl, termc2, ..., termcp}, ал Mdict:=Mdict\{termc1, termc2, ..., termcp} және Mnondlct:-Mnondlct\ {termcl, termc2, ..., termcp} ; Mcorr:=Mcorr\ {combc} болады.
      2. Егер combcMcorr\Msave, (combcMcomb\_br) (MUMdlctUMnondlct) = және үшін combmMcomb дұрыс болса, combmcombc болғандықтан, онда M:=MU{combc} ; Mcorr:=Mcorr\ { combc} болады.
      3. Егер combcMsave\Mcorr және әрбір termc3combc үшін орындалса, онда не termcjMUMnondict, не 3 termdMdict болғанда, termdCtermcj болғандықтан, сонда M:=MU{termcl, termc2, ..., termcp} болады, ал Mdlct:=Mdlct\{termci, termc2, ..., termcp} және Mnondict:=Mnondict\{termc1, termc2, ..., termcp}; Mcorr:=Mcorr\{ combc} болады .

Егер 6-ереже бойынша өңделген comb1Mcombcomb2Mcomb үшін осылай болса, comb1comb2 болғандықтан, онда Mcorr:=Mcorr\ {comb2} болады.

**7-ереже**. Егер 4-6 ережелерді қолданған соң C ≠ | M | болса, онда осы ережелердің қолданысын қайталау керек

8-ереже.G1, G2, ..., GN мәтіндік нұсқалардың топтары жинағын қалыптастырғанда М жиынтығындағы термин-дублеттерді іздеу M(Mcand=M) элементтерінің өз арасында жүзеге асады, ал М-дегі терминдердің лексикалық-синтаксистік нұсқаларын іздеу - M және Mnondlct (Mcand=MUMnondict) жиынтығындағы элементтер арасында болады. Ережені қолдану сооңында Mnondict:Мnondict(G1 G2  ... GN) болады.

Mvar= G1 G2  ... GN- мәтіндік нұсқалар жиынтығы.

9-ереже.С : = |М| болса, одан соң M : =MMvar. Егер C ≠ | М| болса, онда 4-8 ережелерін қайталаймыз.

10-ережені қолдану алдында С : = | М| болады.

10-ереже. Әрбір termn үшін EMnondict болады:

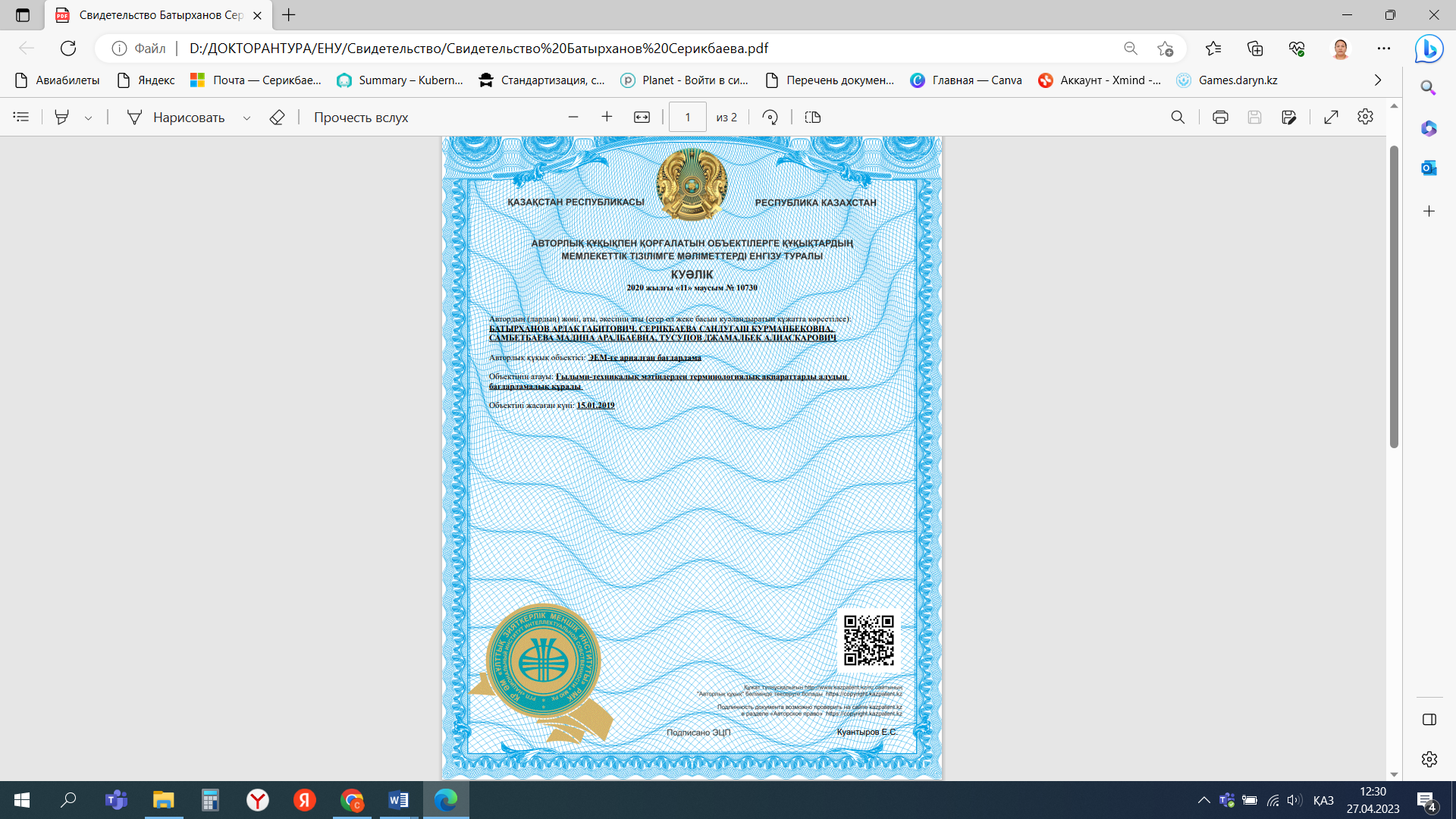
* + - * 1. Егер termn∉Mcomb және |termn|≥[F] болса,онда М:= M{termn} және Мnondict:=Мnondict\ {termn}.
        2. Егер termnMcomb және | termn | ≥ [F]+1 болса, онда M:=M{termn}, ал Мnondict:= Мnondict \ {termn} және Мcomb:=Мcomb\ {termn} болады.

11-ереже. Егер 10-ережені қолданған соң C ≠ | М | болса, онда 4-9 ережелерді қайта қолдану қажет.

12-ереже. Айталық С: = |М|, одан соң М:=MMdict; болады. Егер С: ≠ |М| болса, онда 4-9 ережелерді қайталаймыз.

# **ҚОСЫМША Ғ**

Авторлық куәлік



# **ҚОСЫМША Д**

Ендіру актісі

