



ҚАЗАҚСТАН РЕСПУБЛИКАСЫНЫҢ МЕМЛЕКЕТТІК СТАНДАРТЫ

АҚПАРАТТЫҚ ТЕХНОЛОГИЯ

ДЕРЕКТЕР БАЗАСЫН СЕРТИФИКАТТАУ МӘТІНДІК ДЕРЕКТЕРДЕГІ ЕМЛЕЛІК ҚАТЕЛЕРДІ ТАБУ ӘДІСТЕРІ

ҚР СТ 34.001-2002

Ресми басылым

**Қазақстан Республикасы индустрия және сауда министрлігінің
Стандарттау, метрология және сертификаттау жөніндегі комитеті
(Мемстандарт)**

Астана

Алғысөз

1 ТК 34 «Ақпараттық технологиялар» стандарттау жөніндегі техникалық комитет **ӘЗІРЛЕП ЕНГІЗДІ**

2 Қазақстан Республикасы индустрия және сауда министрлігі Стандарттау, метрология және сертификаттау жөніндегі комитетінің 2002 жылғы 29 желтоқсанда № 519 бұйрығымен **БЕКІТІЛІП ҚОЛДАНЫСҚА ЕНГІЗІЛДІ**

**3 БІРІНШІ ТЕКСЕРУ МЕРЗІМІ
ТЕКСЕРУ КЕЗЕНДІЛІГІ**

**2007 жыл
5 жыл**

4. АЛҒАШ РЕТ ЕНГІЗІЛДІ

Мазмұны

1 Қолданылу саласы	1
2 Жалпы ережелер	1
3 Емделік қателерді табу әдістері	2
4 Сынақ жүргізу рәсімі	5
А қосымшасы (анықтамалық). ДБ сынағы хаттамасының нысаны	9

ҚАЗАҚСТАН РЕСПУБЛИКАСЫНЫҢ МЕМЛЕКЕТТІК СТАНДАРТЫ

**АҚПАРАТТЫҚ ТЕХНОЛОГИЯ
ДЕРЕКТЕР БАЗАСЫН СЕРТИФИКАТТАУ
МӘТІНДІК ДЕРЕКТЕРДЕГІ ҚАТЕЛЕРДІ ТАБУ ӘДІСТЕРІ**

Енгізілген күні 2004-01-01

1 Қолданылу саласы

Осы стандарт деректер базасына (ДБ) ондағы мәтіндік деректердің дұрыстығын бағалау мақсатында сынақ жүргізу әдісін, рәсімін және тәртібін анықтайтын ұйымдық құжат болып табылады.

Стандарт мәтіндік ақпарат ұстайтын ДБ барлық түрлеріне, оның ішінде ДБ-ның толық мәтіндік, реферативтік және лексикографиялық түрлеріне, сондай-ақ объектілердің мәтіндік суреттемелерін және олардың сипаттамаларын ұстайтын фактографиялық деректер базаларына қолданылады.

Стандарт ДБ-на сынақ жүргізген кезде аккредиттелген сынақ зертханаларының персоналдарына қолданылу үшін ұсынылады. Сондай-ақ, оны сынақ зертханасының қызметіне бақылауды жүзеге асыратын мамандар және осы қызметтің нәтижесін қарап отыратын арбитраж органдар да пайдалануы мүмкін. Нақты ДБ-ның сынақтары кезінде осы стандартты пайдалану тәртібі Сынақ бағдарламасына сәйкес белгіленеді.

2 Жалпы ережелер

Деректер базасындағы мәтіндік деректердің дұрыстығы мәтіндік өрістердегі қателердің санымен бағаланады. Деректерді кез келген бұрмалау оның сипатына немесе оның пайда болуына қарамастан қате деп саналады. Қате кездейсоқ (техникалық немесе бағдарламалық құралдардың жаңылысуы, оператордың қателесуі) немесе әдейі (деректерді қасақана, оның ішінде түсінбей бұрмалау) болуы мүмкін.

Осы стандартта тек нақты формальді белгілері бар қателер ғана қаралады. Бұған синтаксистік (меңгеруі немесе қиысуы дұрыс емес) немесе мағыналық қателер (сөздерді дұрыс қолданбау, нақты қателер) емес тек қана емлелік қателер (қолданылмайтын сөзтұлғалар, символдардың жол берілмейтін тіркестері) жатады.

Табиғи тілде мәтінді кез келген бұрмалау емлелік қате деп аталады, нәтижесінде қолданылмайтын сөзтұлға, символдардың (әріптердің немесе тыныс белгілердің) жол берілмейтін тіркестері пайда болады.

Ресми басылым

ДБ-ның мәтіндік өрісінде кездесетін емлелік қателер мынадай типтерге бөлінеді:

- 1) әріптерді қалдырып кету, енгізу, ауыстыру немесе орнын ауыстыру;
- 2) сөздердің арасында бос орын немесе сөздердің ішінде дефисті қалдырып кету немесе, керісінше, сөздердің ішінде артық бос орын немесе дефис қою;
- 3) бір сөзде орыс және латын әріптерін араластыру, сондай-ақ әріптер мен цифрды араластыру;
- 4) бас әріптер мен кіші әріптерді дұрыс қолданбау (сөйлемнің басында кіші әріптерді қолдану және т.б.);
- 5) сөздердің қайталануы;
- 6) тыныс белгілерді дұрыс қолданбау (тыныс белгілерінің жол берілмейтін тіркесімі, тыныс белгісінен кейін бос орынның жоқтығы және тыныс белгісінің алдынан артық бос орындар, абзацтың соңында нүктенің жоқтығы және т.б.).

Жоғарыда айтылған қателердің барлық типтерінің нақты формальді белгілері болады және емлені тексеретін арнаулы бағдарламаның көмегімен автоматты түрде табылуы мүмкін.

Қателердің нәтижесінде бір сөзтұлғаның басқа сөзтұлғаға ауысуы болғанда, мұндай қатені синтаксистік (қиысудың, меңгерудің бұзылуы) немесе мағыналық (сөздердің ауысуы) қателерден ажыратуға мүмкіндік беретін формальді өлшем болмағандықтан қаралмайды.

Мәтіндік емес ақпаратты бұрмалау (цифрлардағы, формулалардағы, шартты белгілердегі қателер) емлелік қателерге жатпайды, сондықтан ол да қаралмайды.

Хаттама нақты ДБ-сын сынаудың нәтижесі болып табылады, онда мынадай деректер (әрбір өріс бойынша жеке-жеке): талданылған мәтіндік деректердің жалпы ұзындығы (белгілерде және сөздерде); типтері бойынша бөлінген айқындалған емлелік қателердің тізімі (егер ол өте көп болса, онда ықтимал типтік шамасы); айқындалған қателердің жалпы саны; қателердің орташа проценті; сондай-ақ бүкіл мәтіндік өрістер бойынша қорытынды деректері: ДБ-сындағы қателердің жалпы саны және қателердің орташа проценті көрсетілген.

3 Емлелік қателерді табу әдістері

3.1. Қателерді табудың сөздік әдісі

Қателерді табудың сөздік әдісі - сөзтұлғалардың барлығын тексеру үшін сөзтүрлендіру туралы мәліметтері, яғни жалғаулардың тізімдері және олардың тіркесім ережелері қоса берілетін өзінде лексиканың көрнекілік бөлшегі бар сөздікті пайдаланудан тұрады. Бұл әдіс нәтижелердің жоғары нақтылығын және сенімділігін қамтамасыз етеді, алайда бұл көп шығынды

(сөздікті құрастыруға және оны күн талабына сай келтіруге, сондай-ақ сөз түрлендіру ережесінің формализациясына), дискілік (сақтау үшін) және жедел жадтың (тез іздеу үшін) үлкен көлемін, іздеудің айтарлықтай уақытын талап етеді.

Іс жүзінде бұл әдісті іске асыру үшін емле тексеретін кез келген қолданылып жүрген бағдарламаны пайдалануға болады. Бұл бағдарламалар сөздіктің көлемімен және тақырыптамалық қамтылуымен, табылған қателер типтерінің санымен, түзету нұсқаларын іздеу жылдамдығымен және нақтылығымен, мәтінді талдау кезіндегі «интеллектуалдық» деңгейімен, пайдаланылатын сөздіктерді құрастыру мүмкіндігімен және т.б. ерекшеленеді. Тексерудің нақты бағдарламасын таңдау сынақ зертханаларының қарап шешуіне қалдырылады.

Тексеру бағдарламасының көпшілігі диалогтік режимде жұмыс істейді және пайдаланушының белсенді қатысуын талап етеді. Тексеру процесінде олар пайдаланушыға табылған қателерді көрсетеді, түзету нұсқаларын және әрекеттердің мүмкін нұсқаларын ұсынады: 1) қатені ескермеу; 2) ұсынылған нұсқалардың біреуімен ауыстыру; 3) сөзді қолмен түзету; 4) қате сөзді белгілеу; 5) сөздікке үстемелеп қосу (негізгі немесе пайдаланылатын) және т.б. Тек кейбір бағдарламалар барлық қате сөздерді автоматты түрде белгілей отырып, дестелік режимде жұмыс істеуі мүмкін.

Емлені тексеру бағдарламасы мынадай талаптарды қанағаттандыруы тиіс:

1) бағдарлама тек мәтіндік редактордың ішін ғана емес, дайын мәтіндік файлдарды біржолата пысықтап шығуы тиіс;

2) бағдарлама бүкіл табылған қателерді арнаулы белгімен белгілей отырып, дестелік режимді қолдап отыруы тиіс;

3) бағдарлама сөздігі қарапайым жалпы ғылыми және публицистикалық мәтіндерге бағытталуы тиіс, сонымен бірге талап етілетін пәндік салаға икемделу мүмкіндігі және пайдаланылатын сөздіктің толығы мүмкіндігі болуы тиіс;

4) бағдарлама сөздігінде базалық лексикадан басқа өзге де нысандардың айрықша көп таралған сөздері болуы тиіс; оның ішінде:

а) жеке есімдер және географиялық атаулар, және де бағдарлама бұл сөздердің барлығы үлкен әріппен жазылатынын ескеруі тиіс;

б) қысқартулар (мысалы, км) және аббревиатуралар (ҚР, СІМ);

в) айрықша көп таралған арнаулы терминдер.

5) бағдарламаның қолданыстағы сөздігінің толығы мүмкіндігі көзделуі тиіс;

6) бағдарлама дефисті сөздерді дұрыс өңдеуі тиіс (дефисті тасымалдан ажырату);

7) бағдарлама сөздікте жоқ сөздерді: неологизмдерді, окказионализмдерді, элементтермен бөлінетін (авто-, авиа-, кино, теле-,

радио-,) сөздерді, еркін сөз тіркестеріне нақты жақын сөздерді бірдей өңдеуі тиіс.

Мәтінді талдау кезінде «интеллектуалдықтың» жоғары деңгейін меңгеретін емлелік қателерді ғана емес, бірқатар синтаксистік (меңгеру немесе қиысудың дұрыс еместігі) және стилистикалық (сөздерді дұрыс қолданбау) қателерді де таба алатын бағдарламаларды пайдалану, сондай-ақ сөздікте жоқ, бірақ бар элементтерден оңай құрастырылатын сөздерді дұрыс өңдеу ұсынылады.

3.2. Тіркесу әдісі

Қателерді табу үшін осы тілдегі әріптердің тіркесімі туралы мәліметтер ғана пайдаланылса тіркесу әдісіне жатады. Бұл әдіс іс жүзінде әріп тіркесінің бүкіл заңдық статистикасы 4-6 әріптен көп емес ұзындықтағы бөлікте анықталатынына негізделген, сонымен бірге қаралатын бөліктер ұзындығының бұдан әрі өсуінен көрініс айтарлықтай өзгермейді. Осы тілдегі әріптер тіркесінің толық көрінісін алу үшін іс жүзінде 3-4 әріп ұзындығындағы тіркесті қарау жеткілікті.

Тіркесу әдісі іске асыруға өте қарапайым және жұмысқа тиімді: ол сөздік құрастыруды талап етпейді, кесте тіркесімі, әдетте, жадтың өзінде орналасады, тексерудің өзі өте тез жүзеге асырылады. Бірақ та бұл әдіс сөздік әдіспен салыстырғанда дәл емес және әрқашан да нақты бар қателерді қалдырып кетуі мүмкін: егер қате сөзтұлғаның тіркесім ережесін бұзбаса, оның табылуы мүмкін емес. Алайда, практика көрсетіп отырғандай қателердің көбісі тіркесім ережесінің бұзылуы бойынша тез анықталатын жеткілікті өрескел жаңылыстарды құрайды. Сондықтан тіркесу әдісінің нақты дәлдігі сөздік әдістің нақты дәлдігінен айтарлықтай ерекшеленбейді.

Тіркесу әдісін іске асыру осы тіл үшін әріптер тіркесімінің кестесін құруды талап етеді. Бағдарламалық жағынан алып қарағанда мұндай кесте 3(4) өлшемдік биттік жиымды білдіреді, бұл жерде өлшем саны қаралатын әріптер тізбегінің ұзындығына тең, ал әрбір өлшемдегі элементтердің саны алфавиттегі әріптердің (орыс тілі үшін 32) санына тең. Мұндай кесте оперативті жадта оңай орналасуы мүмкін. Егер сөздердің шекаралары ескерілсе, яғни бастапқы, ортаңғы және соңғы тіркестерді жеке белгілеп отырса, кестені дәлірек жасауға болады.

Тіркесім кестесі көрінеу емлелік қателерді ұстамайтын ұзындығы 10-20 мың белгі мәтіндер жиымын, сондай-ақ тіркесімнің қалыпты ережесін қатаң бұзатын сөздерді – сирек немесе шетел терминдерін және жеке есімдерді (Ыйсон, Кыргызтан), консонантты қысқартуларды және аббревиатураны (және т.б., км, РК, БД) бірыңғай көшіріп алу жолымен қалыптасады. Көп таралған мұндай сөздер осы әдістің көмегімен анықталған қателерді сүзгіден өткізу үшін қосымша құрал ретінде пайдаланылуы мүмкін жоюдың жеке тізіміне енгізілуі мүмкін.

Мәтін көшіріп алған кезде қалыпты түрге келеді, атап айтқанда:

1) барлық бөлгіш-символдар (бос орын, табуляция, жолдың шекарасы) және тыныс белгілері «сөз шекарасы» (бос орын) символына ауысады;

2) барлық тасымалдар жойылады, ал сөздер қалыпты түрге келеді (бұл жерде сөз бөлшегін құрайтын осы дефис нұсқасы көзделуі тиіс);

3) Барлық бас әріптер кіші әріптерге ауысады (тұтастай бас әріптерден тұратын немесе сөйлемнің ортасында бас әріппен басталатын сөздерді қоспағанда);

4) барлық әріп емес символдар ескерілмейді;

Бұдан кейін осылай ерекшеленген кез келген сөздер көшіріп алынады, одан тізбектеліп ұзындығы 3(4) әріп болатын барлық тіркес шығады және тіркесім кестесіне енгізіледі (тиісінше бит белгіленеді). Егер сөздердің шекаралары ескерілсе, онда кез келген тіркес сәйкес кестеге енгізіледі (бастапқы, ортаңғы немесе соңғы).

Алынған кестені мәтіннің басқа көрнекілік жиымында тестілеу және осы әдістің көмегімен табылған қателердің тізімін сөздік әдістің көмегімен алынған қателердің тізімімен салыстыру ұсынылады.

Тіркесім кестесінің көмегімен қателерді іздеу рәсімі көбінесе кестенің өзінің қалыптасу рәсіміне ұқсас. Сонымен бірге кірістік мәтін де қалыпты жағдайда болады, сөз қалыпты түрге келеді, одан тізбектеліп барлық үш таған (дәптерлер) шығады және сәйкес кесте бойынша тексеріледі. Егер осы тіркес кестеде жоқ болса, онда қате белгіленеді (және де бір сөзде қатарынан бірнеше болуы мүмкін емес). Табылған қате себептері көрсетіле отырып (тыйым салынған тіркес көрсетіледі) және мүмкіндігінше контекст көрсетіле отырып, файл-хаттамаға енгізіледі (жалған қателерді анықтау оңай болу үшін).

Тіркесу әдісі дестелік режимде дайын мәтіндік файлдармен жұмыс істеуге тиісті программалаудың кез келген тіліндегі дербес программа түрінде іске асырылуы мүмкін.

4 Сынақ жүргізу рәсімі

4.1. Сынақтың негізгі кезеңдері

Мәтіндік деректердің дұрыстығын бағалау мақсатында ДБ-на сынақ жүргізу рәсімі мынадай жұмыстарды қамтиды:

1) Тексеру объектісін таңдау және қателерді табу әдісі.

2) Объектіні мәтіндік файлға түсіру.

3) Мәтіндік файл қателерді табу бағдарламасымен өңделеді, жұмыстың нәтижесінде барлық қателер арнаулы белгімен белгіленетін файл қалыптасады (немесе қателердің олардың адрестерімен тізімі бар файл).

4) Қателер тізбесі бар файл арнаулы бағдарламамен өңделеді және нәтижесінде қателердің олардың адрестерімен тізбесі және қателердің жалпы статистикасы бар сынақтың файл-хаттамасы қалыптасады.

4.2. Объектіні таңдау және тексеру әдістері

Емлелік қателерді іздеу рәсімі тек осы тілде толық қанды және байланыстырушы мәтіндік ақпаратты ұстайтын өрістерге қолданылады. Бұл өрістердің құрамы нақты ДБ-ның тақырыптамасы мен құрылымына байланысты. Міндетті тексеруге мынадай деректер өрісі жатады:

реферативтік-библиографиялық ДБ үшін – атауы, реферат немесе аннотация, ескертулер (егер бар болса);

лексикографиялық ДБ үшін – бас әріпті сөз, оның түсіндірмесі, ескертулер (егер бар болса);

толық мәтіндік ДБ үшін - атауы, толық мәтін, ескертулер (егер бар болса);

фактографиялық ДБ үшін – объектінің атауы және объектінің жазбаша суреттемесі немесе оның сапалы сипаттамалары бар барлық өрістер.

Керісінше:

1) цифрлік деректер (сандар, күндер, номерлер);

2) сөздікте жоқ және әріптер тіркесімінің қарапайым ережелерін жиі бұзатын кодтар немесе шартты белгілер бар өрістер қаралмауға тиіс.

Тегі және географиялық (әсіресе, сирек кездесетін немесе шетелдік) атаулары бар өрістерді тексерудің қажеті жоқ, өйткені мұнда дұрыс жазылуын белгілеу мүмкін емес.

Тексерілуге жататын өрістердің тізбесі әрбір нақты ДБ үшін дербес анықталады және Сынақтар бағдарламасына тіркеледі.

Қателерді іздеу әдісін таңдау кезінде іске асыру мен қолдану жағынан қарағанда сөздік әдіс нақтырақ, бірақ көп еңбекті қажет ететінін, ал тіркесу әдісі нақты болмаса да, іске асыруда оңай және айтарлықтай тез жұмыс істейтінін негізге алған жөн. Сондықтан тексеріліп отырған мәтіннің типіне, машиналық уақыттың бар ресурстарына және тексерудің қажетті дәлділігіне байланысты осы әдістердің біреуін немесе екеуін бірге пайдалану керек.

4.3 ДБ-нан мәтіндік деректерді түсіру

Емлені тексеретін барлық бағдарламалар тек мәтіндік файлдармен ғана жұмыс істеп, ДБ-ның файлдарымен тікелей жұмыс істемейтіндіктен мәтіндік деректерді ДБ-нан сыртқы файлға түсіру қажеттілігі туындайды.

ДБ-мен бірге қойылатын жүйе типіне және ДБ-ның өз деректер пішіміне байланысты мынадай нұсқалары болуы мүмкін:

1) Жүйенің мәтіндік файлға деректерді түсіру үшін кіріктірме құралдары бар, яғни талап етілетін нысан есептерінің құрылу мүмкіндігі болады. Бұл жағдайда мәтіндік ақпараты бар тек қажетті өрісті қамтитын есеп пішімін құру қажет;

2) жүйенің өзі мұндай құралдарға ие емес, алайда ол есептердің кіріктіріме генераторы бар белгілі бағдарламаның көмегімен құрастырылған. Бұл жағдайда есеп пішімі осы бағдарламаның құралдарымен құрастырылады;

3) ДБ-ның осы пішіміне арналған талап етілетін пішім есептерінің дайын генераторы жоқ, алайда деректер пішімінің өзі белгілі болуы немесе жеңіл анықталуы мүмкін. Бұл жағдайда тікелей ДБ-нан деректерді түсіру үшін арнаулы бағдарлама жазу қажет болады. Бұл бағдарлама (бір рет қолданылуы мүмкін) құжаттанып, сынақ зертханасы бағдарламасының мұрағатына енгізілуі тиіс;

4) деректер пішімі белгісіз немесе деректер, оған тікелей кіруге болмайтындай етіп арнайы шифрленген. Бұл жағдайда ақпарат экранның көшірмесін алу резиденттік бағдарламасының көмегімен тікелей экраннан алынады. Бұл ретте сынақшы әр жазбаны қолмен бетін санап, қарастыруы, экранның қажетті фрагментін таңдауы және оларды мәтіндік файлға жинақтауы тиіс, тек содан кейін тексеру бағдарламасын пайдалануға болады.

Есеп пішімін құру кезінде барлық қызметтік ақпараттар (өрістердің аты, жазбалардың реттік саны, бөлгіштер) алынып тасталуы тиіс. Егер бұл мүмкін болмаса, онда алынған мәтіндік файлды қолмен немесе автоматты түрде (арнаулы бағдарламаның көмегімен) түзету қажет. Емлелік бақылау бағдарламасының енуіне ДБ-нан тек таңдалған мәтіндік деректері бар мәтіндік файлдар түсуі тиіс.

4.4 Қателерді табу бағдарламасын іске қосу

Бұдан әрі мәтіндік файл тексерілген деректерімен қателерді тексеру бағдарламасының енуіне түседі. Егер тексерудің стандартты бағдарламасы пайдаланылса, дұрыс емес сөзтұлғалар түзетілмей арнаулы белгімен белгіленуі үшін оны дестелік режимде іске қосуға болады. Егер бағдарлама мұндай режимде жұмыс істей алмайтын болса, тексеру процесі диалог режимінде жүргізіледі және ол елеулі уақытты талап етеді. Егер бағдарлама қателердің тізімін генераторласа және қателер санағын жинақтаса, тез арада талапқа ең жақын түрде файл-хаттама алу үшін оны пайдаланып қалу керек.

Егер қателерді іздеудің өз бағдарламасы (тіркесу әдісінің көмегімен) пайдаланылса, ол тез арада талап етілген түрдегі файл-хаттаманы генераторлар беруді жасауы мүмкін.

4.5 Тексеру нәтижесін өңдеу

Қателерді табу бағдарламасының жұмысы нәтижесінде контексті көрсетілген, барлық табылған қателері бар файл-тізімі пайда болады. Тексеру әдісінің қай-қайсысы да абсолюттік дәлдікті бермейтіндіктен тексеру бағдарламасы танымаған жалған қателерді – қолданылып жүрген сөзтұлғаларды одан жою үшін сынақшыға алынған қателердің тізімін

өзінің қарап шығуын ұсынады. Егер бағдарлама өзінің сөздігінде жоқ сирек немесе арнаулы сөздерді тани алмаса, сынақшы арнаулы сөздік бойынша өзі тексеруі керек. Егер ол мүмкін болмаса немесе көп еңбекті қажет ететін болса, жалған қателердің болуына көну керек.

Егер тексеру бағдарламасы талап етілетін түрдегі файл-тізбені бере алмаса, барлық керек емес ақпаратты алып тастау және файлды талап етілетін түрге келтіру үшін оны мәтіндік редактор көмегі арқылы немесе арнаулы фильтр-бағдарламасының көмегі арқылы қолмен редакциялау қажет. Мысалы, егер тексеру бағдарламасы қателер тізімін бере алмай, тек қате сөздерді арнайы белгілермен белгілесе, барлық белгіленбеген сөздерді, ал содан кейін белгілердің өздерін алып тастау керек.

Мынадай мәліметтерді алу үшін алынған қателердің файл-тізбесін (әрбір өріс бойынша жеке) өңдеуден өткізу қажет:

1) Белгілердегі және сөздердегі тексерілетін мәтіндік деректердің жалпы ұзындығы (деректерді ДБ-нан түсіру кезінде анықталады);

2) табылған қателердің жалпы саны (белгілер мен сөздерге қатысты).

Бұдан кейін жеке өрістер бойынша деректерді қосып, барлық ДБ бойынша қателердің орта процентін алу қажет.

Қателердің рұқсат етілген нормасы ДБ-ның мақсатына байланысты ерекшеленеді: ресми ақпараты бар ДБ-да қателер мүлдем болмауы тиіс, ал «Ғылыми техникалық ақпараттар» ДБ-да 2-5% қателердің болуына жол берілуі мүмкін.

А қосымшасы
(анықтамалық)

ДБ сынағы хаттамасының нысаны

Хаттама

_____ ж. «__» _____ №_____
деректер базасының сынағы

Осы хаттама _____ ұсынған
(өтініш беруші ұйымның атауы)
_____ деректер
(ДБ аты)
базасы сынағының нәтижесін куәландырады.

1. Сынақ ДБ мәтіндік деректердің емлелік дұрыстығын бағалау және ондағы емлелік қателерді анықтау мақсатында жүргізілді.

2. Сынақты _____ ж. «__» _____ №_____, №_____ тармақ Сынақ бағдарламасының негізінде _____ жылдан _____ жылға дейін _____ сынақ зертханасының _____ (сынақты жүргізу орны) қызметкерлері _____ жүргізді. _____ (сынақты жүргізген адамның аты-жөні)

3. Тексеру объектісі – ДБ-нан мәтіндік деректер (ДБ-нан таңдау) көлемі _____ жазба, _____ Кбайт.

Мынадай деректер өрісі тексерілді: _____
(тексерілген өрістер тізімі)

4. Сынақ _____ ҚР СТ сәйкес _____ техникалық құралдарда _____ (техникалық құралдардың типі, таңбасы және №) аттестатталған _____ әдісін пайдалану арқылы _____ бағдарламасын пайдалану арқылы жүргізілді. _____ (бағдарламаның тізбесі)

5. Сынақ барысында мынадай нәтижелер алынды:

өрістің аты	белгілер саны	сөздердің саны	қателердің саны	% қате
Барлығы:				

6. Алынған нәтижелердің негізінде осы ДБ-да қателердің орташа проценті _____% (1000 белгіге _____ қате) бар.

Сынақты жүргізген адам _____ (аты-жөні)
_____ ж. " ____ "

Сынақтың жетекшісі _____ (аты-жөні)
_____ ж. " ____ "

ӘОЖ

МСЖ 35.240

Түйінді сөздер: емлелік дұрыстықты бағалау, мәтіндік деректер



ГОСУДАРСТВЕННЫЙ СТАНДАРТ РЕСПУБЛИКИ КАЗАХСТАН

**ИНФОРМАЦИОННАЯ ТЕХНОЛОГИЯ.
СЕРТИФИКАЦИЯ БАЗ ДАННЫХ.
МЕТОДЫ ОБНАРУЖЕНИЯ ОРФОГРАФИЧЕСКИХ
ОШИБОК В ТЕКСТОВЫХ ДАННЫХ**

СТ РК 34.001-2002

Издание официальное

**Комитет по стандартизации, метрологии и сертификации
Министерства индустрии и торговли Республики Казахстан
(Госстандарт)**

Астана

Предисловие

1 РАЗРАБОТАН И ВНЕСЕН Техническим комитетом по стандартизации ТК 34 «Информационные технологии»

2 УТВЕРЖДЕН И ВВЕДЕН В ДЕЙСТВИЕ приказом Комитета по стандартизации, метрологии и сертификации Министерства промышленности и торговли Республики Казахстан от 29 декабря 2002 г. № 519

**3 СРОК ПЕРВОЙ ПРОВЕРКИ
ПЕРИОДИЧНОСТЬ ПРОВЕРКИ**

2007 год
5 лет

4 ВВЕДЕН ВПЕРВЫЕ

Содержание

1	Область применения.....	1
2	Общие положения.....	1
3	Методы обнаружения орфографических ошибок.....	3
3.1	Словарный метод обнаружения ошибок.....	3
3.2	Метод сочетаний.....	5
4	Процедура проведения испытаний.....	7
4.1	Основные этапы испытаний.....	7
4.2	Выбор объекта и метод проверки.....	7
4.3	Выгрузка текстовых данных из БД.....	8
4.4	Запуск программы обнаружения ошибок.....	9
4.5	Обработка результатов проверки.....	9
	Приложение А Форма протокола испытаний БД.....	11

ГОСУДАРСТВЕННЫЙ СТАНДАРТ РЕСПУБЛИКИ КАЗАХСТАН

**ИНФОРМАЦИОННАЯ ТЕХНОЛОГИЯ.
СЕРТИФИКАЦИЯ БАЗ ДАННЫХ.
МЕТОДЫ ОБНАРУЖЕНИЯ ОРФОГРАФИЧЕСКИХ
ОШИБОК В ТЕКСТОВЫХ ДАННЫХ**

Дата введения 2004.01.01

1 Область применения

Настоящий стандарт является организационным документом, определяющим методы, процедуры и порядок проведения испытаний баз данных (БД) с целью оценки правильности содержащихся в них текстовых данных.

Стандарт применяется ко всем видам БД, содержащим текстовую информацию, в том числе полнотекстовым, реферативным и лексикографическим БД, а также фактографическим базам данных, содержащих текстовые описания объектов и их характеристики.

Стандарт рекомендуется для применения персоналом аккредитованных испытательных лабораторий при проведении испытаний БД. Он может быть использован также специалистами, осуществляющими контроль за деятельностью испытательной лаборатории и органами арбитража, рассматривающими результаты этой деятельности. Порядок использования настоящего стандарта при испытаниях конкретной БД устанавливается соответствующей Программой испытаний.

2 Общие положения

Правильность текстовых данных в БД оценивается числом ошибок в текстовых полях.

Ошибкой называется любое искажение данных, независимо от его характера или его возникновения. Ошибка может быть случайной

(сбой технических или программных средств, ошибка оператора) или преднамеренной (умышленное искажение данных, в том числе из-за непонимания).

В настоящем стандарте рассматриваются только ошибки, имеющие четкие формальные признаки. Сюда относятся только собственно орфографические ошибки (несуществующие словоформы, недопустимые сочетания символов), но не синтаксические (неправильное управление или согласование) или смысловые (неправильное употребление слов, фактические ошибки).

Орфографической ошибкой называется любое искажение текста на естественном языке, в результате которого появляется несуществующая словоформа или недопустимое сочетание символов (букв или знаков препинания).

Подразделяются следующие типы орфографических ошибок, встречающихся в текстовых полях БД:

- 1) пропуск, вставка, замена или перестановка букв;
- 2) пропуск пробела между словами или дефиса внутри слова, или, наоборот, вставка лишнего пробела или дефиса внутрь слова;
- 3) смешение русских и латинских букв, а также смешение букв и цифр в одном слове;
- 4) неправильное употребление прописных и строчных букв (строчные буквы в начале предложения, и др.);
- 5) повторение слов;
- 6) неправильное употребление знаков препинания (недопустимые сочетания знаков препинания, отсутствие пробела после знака и лишние пробелы перед знаком препинания, отсутствие точки в конце абзаца и др.).

Все рассмотренные выше типы ошибок имеют четкие формальные признаки и могут быть обнаружены автоматически с помощью специальных программ проверки орфографии.

Случаи, когда в результате ошибки происходит замена одной словоформы на другую, здесь не рассматриваются, поскольку не существует формального критерия, позволяющего отличить такую ошибку от синтаксической (нарушение согласования, управления) или смысловой (замена слова).

Искажение нетекстовой информации (ошибки в цифрах, формулах, условных обозначениях) не относятся к орфографическим ошибкам и поэтому также не рассматриваются.

Результатом испытаний конкретной БД является протокол, в котором указаны следующие данные (в отдельности по каждому полю): общая длина проанализированных текстовых данных (в знаках и словах); список выявленных орфографических ошибок (если он слишком велик, то наиболее типичные примеры) с разделением по типам; общее число выявленных ошибок; средний процент ошибок; а также итоговые данные по всем текстовым полям: общее число ошибок и средний процент ошибок в БД.

3 Методы обнаружения орфографических ошибок

3.1 Словарный метод обнаружения ошибок

Словарный метод обнаружения ошибок заключается в том, что для проверки существования словоформы используется словарь, содержащий в себе представительную часть лексики, сопровождаемую сведениями о словоизменении, т.е. списками окончаний и правилами их сочетаемости. Этот метод обеспечивает высокую точность и надежность результатов, однако, требует больших затрат (на создание словаря и его актуализацию, а также на формализацию правил словоизменения), большого объема дисковой (для хранения) и оперативной (для быстрого поиска) памяти, значительного времени поиска.

Для практической реализации этого метода следует использовать любую из имеющихся программ проверки орфографии. Эти программы различаются объемом и тематическим охватом словарей, числом типов обнаруживаемых ошибок, скоростью и точностью поиска вариантов исправления, степенью «интеллектуальности» при анализе текста, возможностью создания пользовательских словарей и т.д. Выбор конкретной программы проверки остается на усмотрение испытательной лаборатории.

Большинство программ проверки работают в диалоговом режиме и требуют активного участия пользователя. В процессе проверки они

предъявляют пользователю обнаруженные ошибки, предлагают варианты исправления и возможные варианты действий:

- игнорировать ошибку;
- заменить на один из предлагаемых вариантов;
- исправить слово вручную;
- пометить ошибочное слово;
- добавить в словарь (основной или пользовательский) и др.

Лишь немногие программы могут работать в пакетном режиме, автоматически отмечая все ошибочные слова.

Программа проверки орфографии должна удовлетворять следующим требованиям:

- программа должна отрабатывать готовые текстовые файлы, а не только изнутри текстового редактора;
- программа должна поддерживать пакетный режим, отмечая все найденные ошибки специальным значком;
- словарь программы должен быть ориентирован на обычные общенаучные и публицистические тексты, при этом должна существовать возможность настройки на требуемую предметную область или возможность пополнения пользовательского словаря;
- словарь программы, помимо базовой лексики, должен содержать наиболее распространенные слова других категорий, в частности:
 - имена собственные и географические названия, причем программа должна учитывать, что эти слова пишутся с большой буквы;
 - сокращения (например км) и аббревиатуры (РК, МИД);
 - наиболее распространенные специальные термины.
- необходимо предусмотреть возможность пополнения пользовательского словаря программы;
- программа должна правильно обрабатывать слова с дефисом (отличать дефис от переноса);
- программа должна адекватно обрабатывать слова, отсутствующие в словаре: неологизмы, окказионализмы, слова с отделяемыми элементами (авто-, авиа-, кино, теле-, радио-), фактически близкие к свободным словосочетаниям.

Рекомендуется использовать программы, обладающие максимальной степенью «интеллектуальности» при анализе текста, которые умеют обнаруживать не только орфографические, но и некоторые синтаксические (неправильное управление или согласование) и

стилистические ошибки (неправильное употребление слов), а также правильно обрабатывать слова, отсутствующие в словаре, но легко конструируемые из имеющихся элементов.

3.2 Метод сочетаний

Метод сочетаний состоит в том, что для обнаружения ошибок используются только сведения о сочетаемости букв в данном языке. Этот метод основывается на том, что практически все статистические закономерности сочетаемости букв выявляются на отрезке длиной не более 4 - 6 букв, причем, дальнейшее увеличение длин рассматриваемых отрезков существенно не меняет картины. Практически достаточно рассмотреть сочетания длиной 3 - 4 буквы (триграммы и тетраграммы), чтобы получить весьма полную картину сочетаемости букв в данном языке.

Метод сочетаний чрезвычайно прост в реализации и эффективен в работе: он не требует составления словаря, таблица сочетаемости обычно размещается прямо в памяти, сама проверка осуществляется очень быстро. Правда, этот метод менее точен по сравнению со словарным и всегда оставляет возможность пропуска реально существующей ошибки: если ошибочная словоформа не нарушает правил сочетаемости, она в принципе не может быть обнаружена. Однако, практика показывает, что большинство ошибок составляют достаточно грубые опечатки, которые легко выявляются по нарушению правил сочетаемости. Поэтому фактическая точность метода сочетаний не сильно отличается от словарного.

Реализация метода сочетаний требует создания таблицы сочетаемости букв для данного языка. С программной точки зрения такая таблица представляет собой 3(4)-мерный битовый массив, в котором число измерений равно длине рассматриваемых цепочек букв, а число элементов в каждом измерении равно числу букв в алфавите (32 для русского языка). Такая таблица легко может разместиться в оперативной памяти. Таблицу можно сделать более точной, если учитывать границы слов, т.е. отдельно фиксировать начальные, срединные и конечные сочетания.

Таблица сочетаемости формируется путем сплошного сканирования массива текстов длиной 10 - 20 тысяч знаков, заведомо не содержащих орфографических ошибок, а также слов, резко нарушаю-

щих нормальные правила сочетаемости - редких или иностранных терминов и собственных имен (Ыйсон, Кыргызстан), консонантных сокращений и аббревиатуру (и т.д., км, РК, БД). Наиболее распространенные из таких слов могут быть вынесены в отдельный список исключений, который может использоваться как дополнительное средство для фильтрации ошибок, выявленных при помощи данного метода.

При сканировании текст приводится к нормализованному виду, а именно:

- все символы-разделители (пробелы, табуляции, границы строк) и знаки пунктуации заменяются на символ «граница слова» (пробел);
- удаляются все переносы, а слова приводятся к нормальному виду (здесь должен быть предусмотрен вариант с настоящим дефисом, составляющим часть слова);
- все заглавные буквы заменяются на строчные (за исключением слов, целиком состоящих из заглавных букв или начинающихся с заглавной буквы в середине предложения);
- все небуквенные символы игнорируются.

Затем каждое выделенное таким образом слово сканируется, из него последовательно извлекаются все сочетания длиной 3(4) буквы и заносятся в таблицу сочетаемости (устанавливается соответствующий бит). Если учитываются границы слов, то каждое сочетание заносится в соответствующую таблицу (начальную, срединную или конечную).

Полученную таблицу рекомендуется протестировать на другом представительном массиве текстов и сравнить перечень ошибок, обнаруживаемых при помощи данного метода, с перечнем ошибок, получаемым при помощи словарного метода.

Процедура поиска ошибок при помощи таблицы сочетаемости во многом аналогична процедуре формирования самой таблицы. При этом входной текст также нормализуется, слова приводятся к нормальному виду, из них последовательно извлекаются все триады (тетрады) и проверяются по соответствующей таблице. Если данного сочетания нет в таблице, то фиксируется ошибка (причем не может быть несколько подряд в одном слове). Обнаруженная ошибка заносится в файл-протокол с указанием причины (указывается запрещен-

ное сочетание) и, по возможности, контекста (чтобы легче было выявить ложные ошибки).

Метод сочетаний может быть реализован в виде автономной программы на любом языке программирования, которая должна работать с готовыми текстовыми файлами в пакетном режиме.

4 Процедура проведения испытаний

4.1 Основные этапы испытаний

Процедура проведения испытаний БД с целью оценки правильности текстовых данных включает в себя следующие работы:

- выбор объекта проверки и метода обнаружения ошибок;
- выгрузка объекта в текстовый файл;
- текстовый файл обрабатывается программой обнаружения ошибок, в результате работы которой формируется файл, где все ошибки отмечаются специальным значком (или файл, содержащий перечень ошибок с их адресами);
- файл с перечнем ошибок обрабатывается специальной программой и в результате формируется файл-протокол испытаний, содержащий перечень ошибок с их адресами и общую статистику ошибок.

4.2 Выбор объекта и методы проверки

Процедура поиска орфографических ошибок применима только к полям, содержащим полноценную и связную текстовую информацию на данном языке. Состав этих полей зависит от тематики и структуры конкретной БД. Обязательной проверке подлежат следующие поля данных:

для реферативно-библиографических БД - название, реферат или аннотация, примечания (если есть);

для лексикографических БД - заглавное слово, его толкования, примечания (если есть);

для полнотекстовых БД - название, полный текст, примечания (если есть);

для фактографических БД - название объекта и все поля, содержащие словесные описания объекта или его качественных характеристик.

Наоборот, не должны рассматриваться поля, содержащие:

- цифровые данные (числа, даты, номера);
- коды или условные обозначения, которые заведомо отсутствуют в словаре и нередко нарушают обычные правила сочетаемости букв.

Не следует проверять поля, содержащие фамилии и географические названия (особенно редкие или иностранные), поскольку здесь невозможно установить правильное написание.

Перечень полей, подлежащих проверке, определяется индивидуально для каждой конкретной БД и фиксируется в Программе испытаний.

Каждое поле БД необходимо проверять отдельно.

При выборе метода поиска ошибок следует исходить из того, что словарный метод более точен, но и более трудоемок с точки зрения реализации и применения, а метод сочетаний менее точен, но проще в реализации и работает значительно быстрее. Поэтому, в зависимости от типа проверяемого текста, имеющихся ресурсов машинного времени и требуемой точности проверки следует использовать один из этих методов или оба вместе.

4.3 Выгрузка текстовых данных из БД

Поскольку все программы проверки орфографии работают только с текстовыми файлами и не работают напрямую с файлами БД, возникает необходимость выгрузки текстовых данных из БД во внешний файл.

В зависимости от типа системы, поставляемой вместе с БД, и формата данных самой БД, возможны следующие варианты:

- система имеет встроенные средства для выгрузки данных в текстовый файл, т.е. возможность создания отчетов требуемой формы. В этом случае нужно построить формат отчета, включающий только необходимые поля, содержащие текстовую информацию;

- система сама не имеет таких средств, однако она создана при помощи известной программы, имеющей встроенный генератор отчетов. В этом случае формат отчета строится средствами этой программы;

- для данного формата БД не существует готового генератора отчетов требуемой формы, однако сам формат данных известен или может быть легко выяснен. В этом случае потребуется написать специальную программу для выгрузки данных непосредственно из БД.

Эта программа (возможно, одноразового применения) должна быть документирована и включена в архив программ испытательной лаборатории;

- формат данных неизвестен или данные специально зашифрованы так, что прямой доступ к ним оказывается невозможным. В этом случае информация берется непосредственно с экрана при помощи резидентной программы снятия копии экрана. При этом испытатель должен вручную пролистать и просмотреть каждую запись, выбрать нужные фрагменты экрана и собрать их в текстовый файл, только после этого можно использовать программу проверки.

При создании формата отчета следует исключить всю служебную информацию (имена полей, номера записей, разделители). Если это невозможно, то полученный текстовый файл нужно будет отредактировать вручную или автоматически (с помощью специальной программы). На вход программы орфографического контроля должен поступить текстовый файл, содержащий только выбранные текстовые данные из БД.

4.4 Запуск программы обнаружения ошибок

Далее текстовый файл с проверенными данными поступает на вход программы проверки ошибок. Если используется стандартная программа проверки, ее желательно запустить в пакетном режиме, чтобы неправильные словоформы не исправлялись, а помечались специальным значком. Если программа не умеет работать в таком режиме, процесс проверки будет происходить в режиме диалога и потребует существенно больше времени. Если программа умеет генерировать перечень ошибок и собирать статистику ошибок, этим необходимо воспользоваться, чтобы сразу получить файл-протокол в виде, максимально близком к желаемому.

Если используется собственная программа поиска ошибок (при помощи метода сочетаний), можно сделать, чтобы она сразу генерировала файл-протокол требуемого вида.

4.5 Обработка результатов проверки

В результате работы программы обнаружения ошибок должен получиться файл-перечень, содержащий все обнаруженные ошибки, желательно, с указанием контекста. Поскольку никакой метод проверки не дает абсолютной точности, испытателю рекомендуется са-

мому просмотреть полученный список ошибок, чтобы исключить оттуда ложные ошибки - существующие словоформы, почему-либо не опознанные программой проверки. Если программа не опознала редкие или специальные слова, отсутствующие в ее словаре, испытателю рекомендуется самому проверить по специальным словарям, существуют ли такие слова, или обратиться к специалисту. Если же это невозможно или слишком трудоемко, придется смириться с существованием ложных ошибок.

Если программа проверки не может выдать файл-перечень требуемого вида, его потребуется отредактировать вручную при помощи текстового редактора или при помощи специальной программы-фильтра, чтобы удалить всю ненужную информацию и привести файл к требуемому виду. Например, если программа проверки не выдает список ошибок, а просто помечает ошибочные слова специальным значком, нужно будет удалить все непомяченные слова, а затем убрать сами пометы.

Полученный файл-перечень ошибок необходимо подвергнуть статистической обработке, чтобы получить следующие данные (отдельно по каждому полю):

- общая длина проверяемых текстовых данных в знаках и словах (определяется при выгрузке данных из БД);
- общее количество обнаруженных ошибок (относительно знаков и слов).

Затем необходимо суммировать данные по отдельным полям и получить средний процент ошибок по всей БД.

Допустимые нормы ошибок различаются в зависимости от назначения БД: в БД, содержащей официальную информацию, ошибок не должно быть вообще, а в БД «Научно-технические информации» можно допустить наличие 2-5% ошибок.

Приложение А
(справочное)

Форма протокола испытаний БД

Протокол

№ _____ от « ____ » _____ г.
испытаний базы данных

Настоящий протокол удостоверяет результаты испытаний базы данных _____,

(наименование БД)

представленной _____

(наименование организации-заявителя)

1. Испытания проводились с целью оценки орфографической правильности текстовых данных в БД и выявления орфографических ошибок в них.

2. Испытания проводились с _____ г. по _____ г.
в _____ сотрудником испытатель-

(место проведения испытаний)

ной лаборатории _____ на основании

(Ф.И.О. лица, проводившего испытания)

Программы испытаний № _____ от « ____ » _____ г., пункт № _____

3. Объект проверки – текстовые данные из БД (выборка из БД) объемом _____ записей, _____ Кбайт.

Проверялись следующие поля данных:

(список проверенных полей)

4. Проверка проводилась в соответствии с СТ РК _____
с использованием _____ метода на аттестованных
технических средствах _____

(тип, марка и № технических средств)

с использованием программ _____

(перечень программ)

5. В ходе испытаний были получены следующие результаты:

Имя поля	Число знаков	Число слов	Число ошибок	% ошибок
Итого:				

6. На основании полученных результатов можно сделать вывод, что данная БД имеет средний процент ошибок _____% (_____ошибок на 1000 знаков).

Лицо, проводившее испытания _____ (Фамилия, инициалы)
« _____ » _____ г.

Руководитель испытаний _____ (Фамилия, инициалы)
« _____ » _____ г.

УДК

МКС 35.240

Ключевые слова: оценка орфографической правильности, текстовые данные
