



ҚАЗАҚСТАН РЕСПУБЛИКАСЫ  
БІЛІМ ЖӘНЕ ҒЫЛЫМ МИНИСТРЛІГІ  
МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ  
РЕСПУБЛИКИ КАЗАХСТАН  
MINISTRY OF EDUCATION AND SCIENCE  
OF THE REPUBLIC OF KAZAKHSTAN



Л. Н. ГУМИЛЕВ АТЫНДАҒЫ  
ЕУРАЗИЯ ҰЛТТЫҚ УНИВЕРСИТЕТІ  
ЕВРАЗИЙСКИЙ НАЦИОНАЛЬНЫЙ  
УНИВЕРСИТЕТ ИМ. Л. Н. ГУМИЛЕВА  
GUMILYOV EURASIAN  
NATIONAL UNIVERSITY



Студенттер мен жас ғалымдардың  
«Ғылым және білім - 2015»  
атты X Халықаралық ғылыми конференциясының  
БАЯНДАМАЛАР ЖИНАҒЫ

---

СБОРНИК МАТЕРИАЛОВ  
X Международной научной конференции  
студентов и молодых ученых  
«Наука и образование - 2015»

---

PROCEEDINGS  
of the X International Scientific Conference  
for students and young scholars  
«Science and education - 2015»

**УДК 001:37.0**  
**ББК72+74.04**  
**Ғ 96**

Ғ96

«Ғылым және білім – 2015» атты студенттер мен жас ғалымдардың X Халық. ғыл. конф. = X Межд. науч. конф. студентов и молодых ученых «Наука и образование - 2015» = The X International Scientific Conference for students and young scholars «Science and education - 2015». – Астана: <http://www.eni.kz/ru/nauka/nauka-i-obrazovanie-2015/>, 2015. – 7419 стр. қазақша, орысша, ағылшынша.

ISBN 978-9965-31-695-1

Жинаққа студенттердің, магистранттардың, докторанттардың және жас ғалымдардың жаратылыстану-техникалық және гуманитарлық ғылымдардың өзекті мәселелері бойынша баяндамалары енгізілген.

The proceedings are the papers of students, undergraduates, doctoral students and young researchers on topical issues of natural and technical sciences and humanities.

В сборник вошли доклады студентов, магистрантов, докторантов и молодых ученых по актуальным вопросам естественно-технических и гуманитарных наук.

УДК 001:37.0  
ББК 72+74.04

ISBN 978-9965-31-695-1

©Л.Н. Гумилев атындағы Еуразия  
ұлттық университеті, 2015

3. Классификатор УДК - <http://teacode.com/online/udc/>
4. Классификатор УДК – Справочник УДК <http://www.rkts.ru/stans/UDC.pdf>
5. Классификаторы отраслей знания - <http://scs.viniti.ru/rubtree>

УДК 004.81

## **КОРПУСТАҒЫ МӘЛІМЕТТЕРДІҢ ДҰРЫСТЫҒЫН ТЕКСЕРЕТІН БАҒДАРЛАМАНЫҢ МОДЕЛІ**

**Дуйсенова Меруэрт Адирбаевна**  
**Жарас Ергеш Жантуғанұлы**  
[flower013@bk.ru](mailto:flower013@bk.ru)

Л.Н.Гумилев атындағы ЕҰУ «6М060200 - Информатика» мамандығының 2 курс  
магистранттары, Астана, Қазақстан  
Ғылыми жетекші – Г.Т.Бекманова

Акустикалық корпус – программалық құрылғылар арқылы қамтамасыз етіліп коопустың жеке элементтеріне қолжетімділікке ие болатын дыбыстық фрагменттерінің құрылымдалған жиыны. Дыбыстық фрагмент корпусының негізгі бірлігі ретінде дыбыстық сигналдың сандармен белгіленген фрагментін сипаттайды, ол белгілі бір типтегі ассоцирленген ақпаратпен жабдықталған.

Акустикалық мәліметтерді белгілеу және транскрипциялау кезінде оператордың жұмыс жасау барысында қателер саны көбейіп, соның салдарынан осы мәліметтер арқылы үйретілетін акустикалық модельдің жұмыс сапасының төмендеуіне алып келуі мүмкін. Осы жағдайда алып келмеуі үшін бірінші ретте Wavesurfer және Audacity бағдарламаларының көмегімен диктормен дыбысталған сөйлемдерді қайта тыңдаудан өткізіп, жіберілген қателер мен техникалық ауытқуларды белгіленіп отырды. Тексеру барысында көп кездесетін қателердің қатарына мыналарды жатқызуға болады:

- сәйкессіздік – диктор сөйлемді жазба қосылмай тұрып немесе жазбаны сөйлемді аяқтағаннан бұрын өшіріп қояу;
- бұрмалау – диктор мәтінде жазылған сөздердің орнына басқа сөзді айтуы немесе белгілі бір сөзді қалдырып кетуі немесе бұрмалап айтуы;
- шу – сөйлемді тыңдау барысында басқа шулар, терең дем алу, сөйлесулер, жөтелу немесе кітап, журнал парақтарын ашқан дыбыстардың естілуі;
- транскрипциялау нәтижесінде пайда болған сәйкес файлдардың жетіспеушілігі - транскрипциялау барысында сөздердің орфографиялық жазылуы бойынша алынған WRD және орфоэпиялық айтылуы бойынша алынған әрбір WRD файына сәйкес PHN файлдардың жетіспеушілігі. Бұл қатенің болуының себебі белгілеулерді дұрыс жасамауы немесе дұрыс сақталмауы салдарынан.

Белгілеу және транскрипциялау бойынша техникалық жұмыстарды тексеру алдыңғы кезеңде құрылған транскрипциялық файлдары бар Wavesurfer программалық қамтамасы мен Audacity пакетімен бірге қосымша бақылау құрал-жабдығы ретінде пайдаланылды.

Тексеру барысында мына жағдайлар ескерілді:

- сөйлемдер, сөз және дыбыстардың шекараларын қою белгілері;
- фонетикалық транскрипция символын қолданудың нақтылығы.

Техникалық жұмысты тексеру барысында есту қабілеті ғана емес (белгі арасындағы дыбысты есту), сонымен қатар міндетті түрде Waveform графикасы мен Audacity бағдарламасындағы Spectrogram графикасымен салыстырып, дыбыстар талдаудан өткізілді. График толқынын (Waveform) спектр графикасымен (Spectrogram) салыстыру арқылы дыбыс аралықтары анықталды. Кейбір дыбыстардың ара қашықтық шекаралары толқын графигінде жақсы көрінген (мысалы, дауысты және дауыссыз дыбыстар аралығы). Кейбір нюанстар спектр графигінде жақсы көрінген (мысалы, қатаң дауыссыздардың ара қашықтығы).

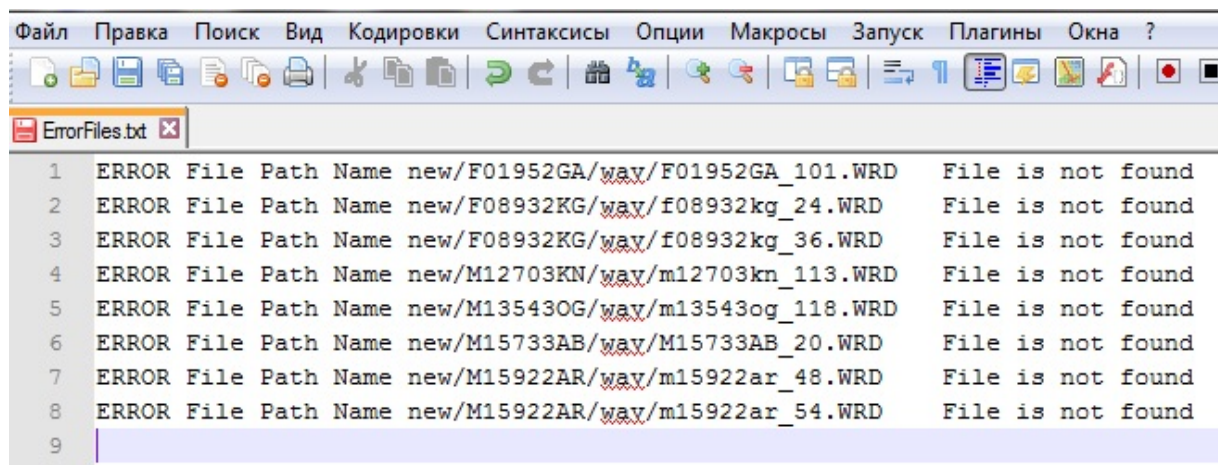
Wavesurfer бағдарламасында мұндай жағдайлардың шешімін табу нашар қарастырылған, сондықтан бұл программада белгілеулер ғана жүргізілді. Шекараларын анықтауда күрделі жағдайлар қатарына мыналарды атап өтуге болады:

- 1) [п, к, қ, т] фонемдері үшін дыбыстық және негізгі тоқтауларын дұрыс анықтау [1];
- 2) дауысты дыбыстар (а, ә, е, и, о, ө, у, ұ, ү, ы, э) және үнді (й, м, н, ң, р) дауыссыздардың аралығын анықтау [1].

Дыбыстар бір-біріне өзара әсер етеді (көрші тұрған дыбыстарға) және дыбыстың жеке бөлігін тыңдау кезінде тұтас тұрған сөздің дыбысталуы кезінде естілетін дыбыстан бөгде дыбысты естуіміз мүмкін. Акустикалық корпустың белгілеуін және транскрипциялануын нақтылау үшін мына процесстерден өткізуіміз керек:

- сөзді тыңдау (алғашында тұтас тыңдау керек, қандай дыбыстардан тұратынын анықтау алу үшін);
- келесі белгілеулерді қою үшін шамамен дыбыс шекараларын анықтау;
- Waveform графигіне қарау;
- Spectrogram графигімен салыстыру;
- белгілеудің орналасуын нақты анықтау;
- сөздің басынан бір бөлігін тыңдау (сөйлемді және шамамен анықталған белгіге дейін);
- шамамен қойылған белгі бөлігі мен сөздің (немесе сөйлемнің) соңына дейін тыңдап, белгіні қою.

Акустикалық корпусты механикалық тексеру барысында адам ресурстары мен уақыт қорын біршама қажет етеді. Корпустың жұмыс өнімділігі мен сапасын арттыру мақсатында тексеру жұмыстарын автоматтандыру мақсатында бағдарламаның моделі жасалынып, жүзеге асырылуда. Бағдарламалық қамтама WRD және соған сәйкес PHN файлдарын тексеріп, екі файлдың бірі жоқ болған жағдайда, жетіспейтін файл атауларын шығарып береді.



1 сурет. Файлдарды тексеру

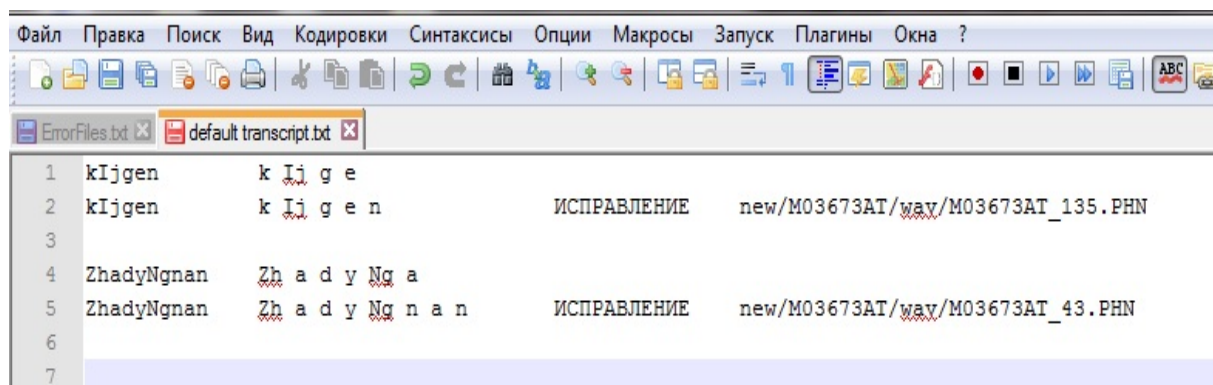
1 суретте көрсетілген ақпаратқа түсіндірме:

- ERROR File Path Name: файлдың қатесі бар туралы ақпарат;
- new/F01952GA/wav: файлдың жолы;
- F01952GA\_101: бағдарламаның тексеру барысында табылмаған файл атауы;
- File is not found: файлды таппағаны туралы қатенің шығуы.

Бағдарламаның жұмыс жасау функциясы: алдымен мәтінді және соған сәйкес дыбыстық файлды тауып алады. Қазақ тілінің орфоэпиялық заңдылықтарына сәйкес мәтін орфографиялық нұсқадан орфоэпиялық нұсқаға қарай ауыстырылады. Мұнда қазақ тілінің орфоэпиялық сөздіктері де ескеріліп дұрыс болуы мүмкін бірнеше нұсқаларын шығарып береді. Содан кейін мәтіннің орфоэпиялық нұсқасын қолдану арқылы берілген дыбыстық файлдың шамамен белгілеу нұсқасы құрылады. Дыбыстың акустикалық қасиеттеріне сәйкес әрбір фонетикалық бірліктің өзінің дыбысталу ұзақтығы болады. Осыған сәйкес, фонетикалық дыбыстың әрбір бірлігіне белгілеу үшін бағдарламада критеріі ретінде береміз.

Мысалы, [a] дыбысының айтылу ұзақтығының орташа мәні 40 мили секундтан жоғары және 130 мили секундтан төмен [2] (кейбір жағдайда дауысты дыбыстарды жасанды түрде ұзаруы мүмкін) болады.

Нәтижесінде әрбір мәтіннің және оның дыбыстық файлының белгілеу критеріімен бірге моделі шығарылады. Егер бағдарлама тексеру барысында белгіленген файл табылса ол кем дегенде бір генерацияланған модельдің белгілеу критерілерімен сәйкестігін тексеріледі. Егер бір модельге болсын сәйкестігі табылса белгілеу файлдары дұрыс болады. Ондай болмаған жағдайда бағдарлама сәйкессіздігі аз белгілеу моделінің бірін тауып береді. Бағдарлама дұрыс белгіленбеген файл мен сәйкессіздік сілтемесі мен файлды жөндеудің мүмкін жолының бірін шығарып береді.



2 сурет. Корпуста сөздердің дұрыстығын тексеру

Мұнда 2 суретте:

- KIJge: сөздің қате екенін көрсетуі;
- KIJgen: жөндеудің мүмкін нұсқасы;
- new/M03673AT/wav/M03673AT\_135.PHN – файлдың жолы.

Бағдарлама C# объектіге бағытталған тілінде жүзеге асырылуда. Корпустағы мәліметтердің дұрыстығын тексеруді автоматтандыру транскрипциялау нәтижесінде жіберілген қателерді жөндеп, механикалық жұмысқа кететін уақыт шығынын азайтуға мүмкіндік береді.

#### Қолданылған әдебиет

1. Шарипбай А. «Создание акустического корпуса казахского языка и уточнение его фонетического строя, представление казахских фонем в международном фонетическом алфавите». Астана, 2012-2014 жоба.
2. Дуйсенова М.А., Бекманова Г.Т. Қазақ тілінің дауысты дыбыстарына акустикалық талдау жасау. /IV Международная научно-практическая конференция «Информатизация общества 2014», Астана 2014.

ӘОЖ:004.8

#### ҚАЗАҚ ТІЛІНДЕГІ МӘТІНДІ МАШИНАЛЫҚ ОҚЫТУ АЛГОРИТМДЕРІ АРҚЫЛЫ АВТОМАТТЫ ТҮРДЕ ТОПТАСТЫРУ

**Искаков Рауан Оразгалиевич**

Магистрант, Семей қаласының Шәкәрім атындағы мемлекеттік университеті, Семей қ.

**Оспанов ЕрболАмангазович**

Семей қаласының Шәкәрім атындағы мемлекеттік университеті, «Автоматика және электротехника» кафедрасының аға оқытушысы