МРНТИ 20.01.45, 28.21.19
УДК 371.39, 519.72

https://doi.org/10.51889/2022-1.1728-7901.21

А.Б. Турдина[1], А.Л. Семенов[2], А.М. Мубараков[1], А.О. Керимбаев [3]*

[1]Л. Гумилев атындағы Евразиялық ұлттық университеті, Нұр-Сұлтан қ., Қазақстан
[2]М.В. Ломоносов атындағы Мәскеу мемлекеттік университеті, Мәскеу қ., Ресей Федерациясы
[3]С. Сейфуллин атындағы Қазақ агротехникалық университеті, Нұр-Сұлтан қ., Қазақстан
**e-mail: aibek.kerimbayev@kazatu.kz*

АҚПАРАТТАРДЫ ТАСЫМАЛДАУДЫҢ ТИІМДІЛІГІН АРТТЫРУДАҒЫ КОДТАУ ТЕОРИЯСЫНЫҢ РӨЛІ ЖӘНЕ ОРНЫ

Аңдатпа

Ақпарат теориясының негізгі әдістеріне кодтау және декодтау жатады. Мақалада ақпараттарды кодтау әдісі зерттелген. Мәліметтерді қодтау мынандай үш мәселені шешеді: хабарламаларды құпияландыру; ақпаратты тығыздау (сжатие информации); хабарламалардағы кездейсоқ қателерді табу және түзету. Жұмыста кодтау теориясының бір мәселесі ақпараттарды тығыздау процедурасы жан-жақты қарастырылған. Ақпараттарды тығыздауда қолданылған ең алғашқы кодтық таңбалар – Морзе әліппесі қарастырылған; Бодо және Фано кодтарының ерекшеліктері ашылған; екілік санақ жүйесін ақпараттарды кодтауда және өлшеуде қолдану процедурасы қарастырылған. Сонымен қатар, қазақ тіліндегі сөздер мен сөйлемдерді кодтау мүмкіндіктері зерттелген. Қазақ тіліндегі мәтіндер зерттеліп, сөздерді құрайтын әріптер, тыныс белгілері және цифрлардың мәтінде кездесу жиілігі табылған. Қазақ тілінде берілген хабарламалардағы «артық» ақпараттар мөлшері есептеліп, шамасы нақты анықталды.

Түйін сөздер: кодтау теориясы, ақпаратты тығыздау, Морзе әліппесі, Бодо және Фано коды, артық ақпарат.

Аннотация

А.Б. Турдина[1], А.Л. Семенов[2], А.М. Мубараков[1], А.О. Керимбаев [3]
[1]Евразийский национальный университет им. Л.Гумилева, г. Нур-Султан, Казахстан
[2] Московский государственный университет им. М.В.Ломоносова, г.Москва, Российская Федерация
[3]Казахский агротехнический университет им. С.Сейфуллина, г. Нур-Султан, Казахстан

РОЛЬ И МЕСТО ТЕОРИИ КОДИРОВАНИЯ В УЛУЧШЕНИИ ЭФФЕКТИВНОСТИ ПЕРЕДАЧИ ИНФОРМАЦИИ

Основные методы теории информации включают кодирование и декодирование. В статье изучен метод кодирования информации. Кодирование сообщений решает три задачи: засекречивание сообщений; уплотнение информации (сжатие информации); обнаружение и исправление случайных ошибок в сообщениях. В работе подробно рассмотрена одна из задач теории кодирования - процедура уплотнения информации. Рассмотрены самые ранние кодовые символы, используемые при уплотнении информации - азбука Морзе; раскрыты особенности кодов Бодо и Фано; рассмотрена процедура использования двойчной системы исчесления при кодировании и измерении информации. Кроме того, изучены возможности кодирования слов и предложений в казахском языке. Изучены тексты на казахском языке, найдена частота появления в тексте букв, знаков препинания и цифр. Подсчитано и определено количество «избыточной» информации в сообщениях, переданных на казахском языке.

Ключевые слова: теория кодирования, сжатие информации, азбука Морзе, код Бодо и Фано, избыточная информация.

Abstract

THE ROLE AND PLACE OF CODING THEORY IN IMPROVING THE EFFICIENCY OF INFORMATION TRANSMISSION

Turdina A.B.[1], Semenov A.L.[2], Mubarakov A.M.[1], Kerimbayev A. O.[3]
[1]L. Gumilev Eurasian National University, Nur-Sultan, Kazakhstan
[2]Lomonosov Moscow State University, Moscow, Russian Federation
[3]Kazakh Agrotechnical University named after S. Seifullin, Nur-Sultan, Kazakhstan

The main methods of information theory include encoding and decoding. The article examines the method of encoding information. Message encoding solves three tasks: classifying messages; compacting information (compressing information); detecting and correcting random errors in messages. The paper

considers in detail one of the tasks of the coding theory - the procedure for compacting information. The earliest code symbols used in compacting information - Morse code-are considered; the features of the Bodo and Fano codes are revealed; the procedure of using a binary system of calculus in encoding and measuring information is considered. In addition, the possibilities of encoding words and sentences in the Kazakh language were studied. The texts in the Kazakh language were studied, the frequency of occurrence of letters, punctuation marks and numbers in the text was found. The amount of "redundant" information in the messages transmitted in the Kazakh language has been calculated and clearly defined.

Keywords: coding theory, information compression, Morse code, Bodo and Fano code, redundant information.

Кіріспе

Кодтау теориясы ақпараттық теориямен тығыз байланысқан және математика мен компьютерлік ғылымдардың құрамына кіреді. Мәліметтерді қодтаудың негізгі үш бағыты бар: 1) хабарламаларды құпияландыру; 2) ақпаратты тығыздау (сжатие информации); 3) хабарламалардағы кездейсоқ қателерді табу және түзету. Біз бұл мақалада ақпараттарды *тығыздаудағы* кодтау әдістерінің қолданылу аспектілерін қарастырамыз. Ақпаратты тығыздау арқылы хабарламаларды байланыс каналдары арқылы үнемді және тиімді тасымалдауға болады.

Акпараттық энтропияның формуласы ашылған соң оның қолданбалы жағын дамыту басталды [1]. Аталған формула ақпаратты тасымалдауға қажетті кодтарды зерттеу кезінде жан-жақты қолданыс тапты. Кодтау процедурасы мәліметті немесе хабарды байланыс каналдары (телеграф, телефон, радио, теледидар және т.б.) арқылы тасымалдау мәселесін шешу кезінде пайда болды. Мұндағы кодтаудың негізгі мақсаты ақпаратты құпияландыру емес, хабарды тасымалдауды тез, ыңғайлы және сенімді ету еді. Осы мақсатқа арналған код жасайтын құрылғы хабарламадағы әрбір символды, немесе тұтас бір сөз немесе фразаны белгілі бір сигналдар комбинациясына айналдыра алады. Осы сигналдар комбинациясын *код* немесе *кодтық сөз* деп атайды. Ал хабарды сигналдардың тізбегіне айналдыру операциясы *кодтау* (кодирование) деп, ал кері операцияны, яғни қабылданған сигналдарды бастапқы хабар қалпына келтіру операциясын – *декодтау* (декодирование) деп атайды [2-3].

Бұдан байқайтынымыз, мәліметтерді немесе хабарларды әр түрлі кодтық сөздермен таңбалауға болады. Егер олай болмаса кодтық сөздер көмегімен жеткен хабарды қалпына келтіре алмас едік.

Зерттеу әдіснамасы және нәтижелері

Хабарды тасымалдауға арналған ең бірінші код – Морзе әліппесі болып табылады. «Морзе әліппесі» үштік кодқа жатады. Бұл код бойынша әрбір әріпке немесе цифрға паузамен бөлінген электр тогының қысқа (нүкте) және ұзақтау (сызықша) импульстерінің, сигналдар арасындағы ұзын пауза (әріптер арасын бөледі) және екі есе ұзын пауза (сөздер арасын бөледі) өзіндік тізбектері сәйкес келеді.

Кейінірек телеграфта тек импульс пен паузадан тұратын екі элементар сигналдар (Бодо коды) кең түрде қолданыла бастады. Бұл екі сигналдың ұзақтығы бірдей. Бодо коды бойынша әрбір әріпке немесе тыныс белгісіне бес осындай сигналдан тұратын өзіндік кодтық сөз сәйкес келеді. Екі әр түрлі екі элементар сигналдан ғана тұратын код – екілік деп аталады. Мұнда кодтық сөзді тек 0 және 1 символдарымен белгілеуге болады. Екілік кодының мағынасын толық түсіну үшін мынандай мысалды қарастырайық.

Мысал 1. Бір адам 0 және 7 арасындағы бір санды ойласын (Бұл аралықта 8 сан бар). Осы санды табушы екінші адамға «иә» немесе «жоқ» деген сөздерден тұратын ғана жауап алатын сұрақтар қою керек.Ойланған санды тезірек табу үшін сұрақтарды қалайша қою керек.

Ең қарапайым жол – сәттілікке үміт артып, сандардың бәрін кез-келген тәртіппен айтып шығуға болады. Бұл жағдайда егер сәттілік болса – бір сұрақтан кейін ойланған сан табылады, ал егер сәтсіздік тосып тұраса онда 7 сұрақ қойып ғана санды табуға болады. Сәттілікті күтпей сұрақты тиімді түрде беруге тырысайық. Тиімді сұрақтар жүйесі көмегімен «иә» немесе «жоқ» деген жауаптар көмегімен ойланған сан жөнінде көбірек ақпарат алуға тырысайық. Мысалы, бірінші сұрақты былай қойсақ: Ол сан 0 мен 3 тің арасында ма? «Иә» немесе «жоқ» жауаптарының екеуі де бізді мақсатқа бірдей жақындатады. Ол санды табудың 4 мүмкіндігі қалады.

Егер бірінші сұраққа қанағаттанарлық жауап алсақ, мынандай сұрақ қоюға болады: ол сан 0 немесе 1 саны ма? Егер бірінші сұраққа теріс жауап алсақ онда мынандай сұрақ қойылуы мүмкін: ол

сан 4 не 5 пе? Екінші сұрақ қойылғаннан кейін қандай жағдай болсын санды табудың 2 мүмкіндігі қалады. Енді осы санды табу үшін бір сұрақ қойсақ жеткілікті. Сонымен ойланған санды табу үшін ең аз дегенде 3 сұрақ қойсақ жеткілікті. Ал бұдан аз сұрақ жеткіліксіз екенін дәлелдеуге болады.

Егер «иә» немесе «жоқ» деген мүмкін жауаптарды 0 және 1 символдарымен белгілесек, онда сұраққа алынған жауаптар нөл және бірдің тізбегі түрінде жазылады. Мысалы, ойланған сан 0 болса, онда алынатын үш жауап «иә» болады. Үш «иә» 000 тізбегіне сәйкес келеді.

Егер ойланған сан 3 болса, онда жауаптар «иә», «жоқ», «жоқ» болады. Демек 3 санына 011 тізбегі сәйкес келеді. Жауаптардың нәтижелері бойынша төмендегі кестені толтырамыз.

Кесте 1. Белгілі бір мәліметті табу кезіндегі жауаптарды 0 және 1 символдарымен таңбалау

Ойланған сан	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Жауаптар	*000*	*001*	*010*	*011*	*100*	*101*	*110*	*111*

Екінші жағынан, бұл кестеден ондық позициялы санау жүйесі және екілік позициядағы санау жүйесі арасындағы сәйкестік кестесін көріп отырмыз. Сонымен қатар мынаны байқау қиын емес: 0 және 7 арасындағы сандардың жиынының орнына 8 хабардан тұратын жиынды ұзындығы 3 ке тең және 0 мен 1 ден тұратын тізбекпен кодтауға болатыны көрініп тұр. Егер мұнан да ұзын екілік тізбекті пайдалансақ, онда кез-келген хабарлар жиынын кодтауға болады.

Шындығында, ұзындығы 3 ке тең екілік тізбектің саны $2^3=8$ (олардың барлығы таблицада келтірілді), ұзындығы 4 ке тең екілік тізбектің саны екі есе көп $2^4=16$. Сонымен, ұзындығы n ге тең екілік тізбектің саны 2^n. Сол себепті, 125 хабарды 0 және 1 арқылы кодтау керек болса, ол үшін ұзындығы 7 ге тең екілік тізбектері артығымен жетеді (біздің қарамағымызда $2^7 = 128$ бар). Соңғы мысалдан мынаны түсіндік: M хабарды $2^n \geq M$ немесе $n \geq \log_2 M$ болғанда ғана ұзындығы n болатын екілік тізбекпен кодтауға болады.

Хабарды, мәліметті кодтауды терең түсіну үшін мына мысалды қарастырайық.

2 мысал. Белгілі бір мемлекетте автомобильдердің номері 7 символдан тұрады. Бұл символдар бас әріптерден (26 әріп) және кез-келген тәртіппен жазылатын ондық позициядағы цифрдан тұрады. Әрбір символ бірдей және неғұрлым аз битпен кодталған. Ал әрбір нөмір бірдей және неғұрлым аз байтпен кодталған. Осы автомобильдердің нөмірі компьютерге енгізілген. Мемлекетте шамамен 2 миллиондай жеңіл автомобиль бар. Осы автомобильдердің нөмірін сақтауға бір компьютердің жадысының көлемі жеткілікті ме?

1) 26 әріп + 10 цифр = 36 символ қолданылады.

2) 36 вариантты кодтау үшін 6 бит ақпарат керек, өйткені $2^5=32 < 36 \leq 2^6=64$, яғни 5 бит жетпейді (онымен 32 вариантты кодтауға болады).

3) Сонымен әрбір символға 6 бит ақпарат керек (ең аз ақпарат мөлшері)

4) Толық нөмір 7 символдан тұрады, олардың әрқайсысына 6 бит ақпарат қажет, демек бір нөмірге $6\times7=42$ бит ақпарат керек.

5) Тапсырма шарты бойынша әрбір нөмір байттың бүтін сандарымен кодталады (бір байтта 8 бит бар), сондықтан бір нөмерге 6 байт ақпарат мөлшері қажет ($5\times8=40 < 42 \leq 6\times8=48$), ал бес байт жетпейді, ал алты – ең аз ақпарат мөлшері

6) 2 миллион номерге $20\times6=120$ байт $2\times10^6\times6=1{,}2\times10^7$ байт = 12 Мбайт ақпарат мөлшерін компьютер жадысынан бөлу керек.

7) Компьютер жадысының көлемі жеткілікті.

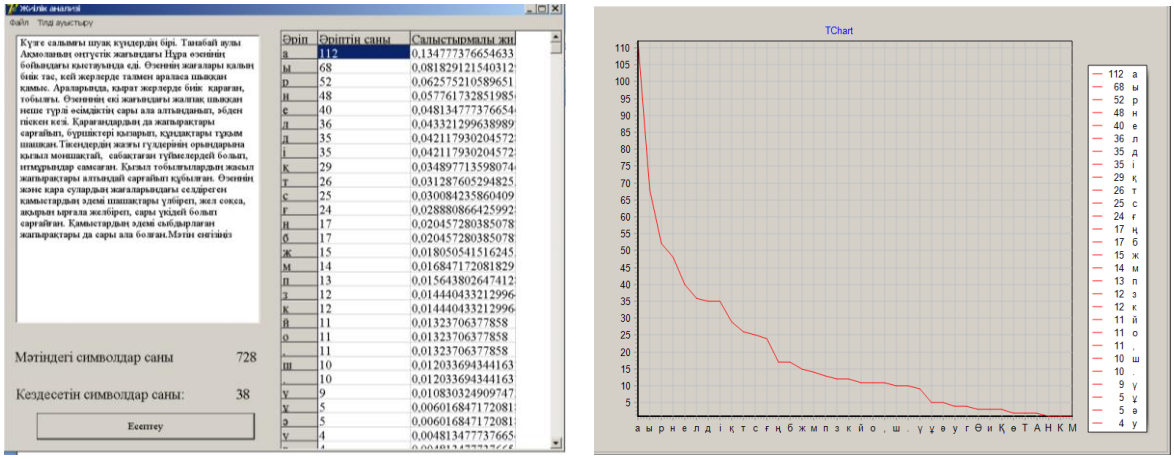
Қазақ тіліндегі хабарламаны кодтау мәселесі. Қазақ тілінде 42 алфавит бар. Сонымен қатар 9 тыныс белгілері және 10 цифр хабар болады. Сонда 61 түрлі элементтер мәтінде кездеседі. Осы элементтердің барлығы тең ықтималдықты деп алсақ, және оларды 0 және 1 символдарымен белгілесек ($2^6 = 64$) ұзындығы 6 ға тең 64 екілік тізбек жеткілікті. Яғни 6 бит ақпарат жеткілікті. Ал мәтін n әріптен тұрса, онда 6 n бит информация кодталған мәтінді тасымалдауға жеткілікті.

Қазақ тіліндегі 61 түрлі элементтің ақпараттық энтропиясы $\log_2 61 = 5{,}9312$ битке тең. Бұл шама 6 биттан аз. Сонымен қатар, қазақ тіліндегі алфавиттердің, тыныс белгілерінің және цифрлардың қолданылу жиілігі де әр түрлі, яғни тең емес ықтималдық жағдайында болады. Мысалы «а» әрпінің мәтінде пайда болу ықтималдығы 0,091 болса, «h» әрпінің пайда болу ықтималдылығы 0,007 ге тең.

Демек, Бодо кодын қолдансақ информацияның байланыс каналымен тасымалдануы үнемді болмауы мүмкін. Бұл жағдайда шеннондық информация ең үлкен шамада болады. Сондықтан, жиілігі үлкен элементтерді (мысалы, «а» әрпі) сипаттайтын кодтың ұзындығы аз, ал сирек пайда болатын элементтердің («h» әрпі) коды ұзын болуға тиіс. Нәтижесінде кодтық мәтін орташа алғанда қысқа болады, ал оны байланыс каналдарымен тасымалдауға аз уақыт кетеді.

Осы бір қарапайым идеяны бірінші рет американдық инженер Морзе қолданған еді. Ол өзінің әліппесін жасау үстінде типографияға барып, дайын әріптердің және тыныс белгілерінің литерлерін санаған. Қай әріптің литерлері ұяшықта көп болса (бұл әріптер мәтінде жиі кездеседі), осы әріп үшін кодты қысқа жасаған, ал сирек кездесетін әріп үшін код ұзын болған. Мысалы, орыс тіліндегі Морзе әліппесінде пайда болу ықтималдығы үлкен «е» әрпі үшін – нүкте, ал «ц» әрпіне төрт символдың тізбегін берген.

Қазақ тілінің әріптерінің мәтіндердегі кездесу ықтималдықтарын табу мақстатында кешенді тәжірибе жасалды. 500-1000 сөзден тұратын бірнеше мәтін жасалды. Әр мәтін қазақ әдебиетінің нақты бір жанрына тиесілі болды. Мынандай жанрлар қамтылды: ертек, әңгіме, повесть немесе роман, өлең және поэма, ғылыми-техникалық мәтін және т.б. Әрбір мәтіндегі әріптердің жиілігін табу процесін автоматтандыру мақсатында Delphi ортасында арнаулы бағдарлама жасалды 1а суретте осы бағдарламаның интерфейсі көрсетілген [4]. Ал сурет 1б да әр әріптің мәтінде кездесу ықтималдылығы арасындағы байланысты сипаттайтын график берілген.



а) *б)*

Сурет 1. Delphi ортасында жасалған бағдарлама және әрбір әріптің және тыныс белгілерінің мәтінде кездесу ықтималдылығының графигі

2 кестеде жасалған кешенді тәжірибенің нәтижесі көрсетілген. Кестеде тек әріптердің мәтінде пайда болу ықтималдылығын ғана есептедік.

Кесте 2. Қазақ тілінің алфавиттерінің мәтіндердегі кездесу ықтималдықтары

№	Әріп	Әріп жиілігі	№	Әріп	Әріп жиілігі	№	Әріп	Әріп жиілігі	№	Әріп	Әріп жиілігі
1	**а**	0,1371	11	**і**	0,0436	21	**з**	0,0166	31	**ю**	0,0018
2	**ы**	0,0783	12	**б**	0,0312	22	**п**	0,0127	32	**х**	0,0056
3	**е**	0,0631	13	**д**	0,0307	23	**ұ**	0,0124	33	**h**	0,0002
4	**т**	0,0616	14	**к**	0,0278	24	**ң**	0,0109	34	**в**	0,0002
5	**л**	0,0543	15	**м**	0,0251	25	**г**	0,0108	35	**ф**	0,0001
6	**и**	0,0500	16	**ш**	0,0250	26	**ү**	0,0107	36	**щ**	0,0001
7	**қ**	0,0488	17	**ғ**	0,0201	27	**ө**	0,0098	37	**э**	0,0001
8	**р**	0,0487	18	**й**	0,0192	28	**и**	0,0090	38	**ь**	0,0000
9	**у**	0,0486	19	**ж**	0,0178	29	**ә**	0,0059	39	**ц**	0,0000
10	**с**	0,0478	20	**о**	0,0166	30	**я**	0,0030	40	**ъ**	0,0000

Тиімді кодтау әдістерін талдау. Сонымен, қазақ мәтініндегі әрбір элементтің (әріп, тыныс белгісі, цифр) пайда болу ықтималдықтарын ескере отырып олардың кодын жасасақ, онда *n* әріптен тұратын мәтінді тасымалдау үшін 6*n* бит ақпарат емес, мысалы 5*n* бит ақпарат қажет болуы мүмкін. Бұл үшін Бодо кодына қарағанда үнемді болатын код жасау керек. Міне осы шарттарды Р.Фано коды орындай алады. Фано коды бойынша байланыс каналдарымен тасымалданатын хабарламаларда оның әрбір элементін оның мәтінде пайда болу ықтималдықтарын ескере отырып жасалған. Нәтижесінде хабардың тасымалдануы тез, ыңғайлы және сенімді болды. Хабардың ақпараттық энтропиясы ең аз шамаға дейін төмендеген [5,6].

Мысал 3. A_1, A_2, A_3, A_4 қазақ тіліндегі хабарламалардың ықтималдықтары $P(A_1) = ½$, $P(A_2) = 1/4$, $P(A_3) = P(A_4) = 1/8$ болсын. Бұл мәліметтерді былай түсіну керек: 1000 хабарламалар жасалған соң 500 рет A_1 хабарлама пайда болады, 250 рет A_2 - хабарлама, ал A_3 және A_4 хабарламалардың әрқайсысы 125 рет пайда болады. Бұл хабарламаларды Бодо әдісімен кодтаймыз:

Кесте 3. Бодо әдісімен ықтималдықтары әр түрлі хабарламаларды 0 және 1 символдарымен таңбалау

A_1	A_2	A_3	A_4
00	01	10	11

Бодо әдісімен кодтау кезінде барлық хабарламаларға бірдей ұзындықтағы код берілген (бірқалыпты код). Бірақ мұндай кодтау кезінде хабарламаның пайда болу ықтималдығы есепке алынбайды. Кодтауды басқаша жасаймыз. Хабарламаларды екі ықтималдықтары бірдей топтарға бөлеміз: бірінші топқа A_1 хабарлама, ал екінші топқа A_2, A_3, A_4 хабарламалар жатады. Бірінші топты 0, ал екіншісін 1 символымен белгілейміз (Кесте 4). Кестенің екінші графасына әр хабарламаның ықтималдықтары енгізілген.

Кесте 4. Хабарламаларды ықтималдықтары бірдей екі топқа бөлу процедурасы

<table>
<tr><td>A_1</td><td>1/2</td><td>0</td><td></td><td></td></tr>
<tr><td>A_2</td><td>1/4</td><td rowspan="3">1</td><td>0</td><td></td></tr>
<tr><td>A_3</td><td>1/8</td><td rowspan="2">1</td><td>0</td></tr>
<tr><td>A_4</td><td>1/8</td><td>1</td></tr>
</table>

Сонымен қатар, A_2, A_3, A_4 хабарламалар да ықтималдықтары бірдей екі топқа бөлінген. Сол сияқты, A_3, A_4 хабарламалар да ықтималдықтары бірдей екі топқа бөлінген. «Хабарлама бірінші топқа жатады ма?» сұрағына 0 символы, немесе «иә» жауабы сәйкес келеді, 1 символы немесе «жоқ» жауабына сәйкес болады.

A_1 хабарламасы талдаудың бірінші адымында ақ «жеке топ» жасап алды. Және оған 0 символы берілді және осы таңба хабарламаның коды болады. Талдаудың екінші адымында A_2 хабарламасы жеке топқа көшті: бірінші адымда хабарлама 1 символына ие болса, екінші адымнан кейін 0 символын алды. Сондықтан бұл хабарламаны 10 таңбасымен кодтаймыз. A_3 және A_4 хабарламалар үшін сәйкесінше 110 және 111 кодтарын береміз. Нәтижесінде төмендегідей кесте жасаймыз.

Кесте 5. Фано әдісімен ықтималдықтары әр түрлі хабарламаларды 0 және 1 символдарымен таңбалау

A_1	A_2	A_3	A_4
0	10	110	111

5 кестеде берілген кодтау әдісін американдық математик Фано ұсынған еді. Фано әдісімен кодтау кезінде барлық хабарламалардың ұзындықтары олардың ықтималдықтарына сәйкес әр түрлі болады (бірқалыпсыз код). Ықтималдылығы үлкен хабарламаның ұзындығы аз, ал ықтималдылығы үлкен боласа ұзын болады. Бодо кодымен салыстырғанда Фано кодының артықшылығы бар. Осы артықшылықты түсіну үшін мынандай мысалды қарастырайық. 1000 хабарды байланыс каналы

арқылы тасымалдау керек. Осы 1000 хабар ішінде A_1 хабар 500 рет, A_2 хабар 250 рет, A_3 және A_4 хабарлар 125 реттен пайда болады. Алдымен ұзындығы екіге тең біркелкі Бодо кодын қолдансақ, онда 2000 екілік символды тасымалдауға тура келеді. Енді Фано кодын қолданамыз. A_1 хабарды бір символмен, A_2 хабарды екі символмен, ал A_3 және A_4 хабарларды 3 символмен кодтаймыз. Сонда жалпы кодтауға жұмсалған символдың саны 500+2·250+(3·125+3·125)=1750 символ. Әркелкі Фано кодын пайдалана отырып хабарды байланыс каналы арқылы тасымалданғанда уақыттың сегізден бір бөлігін үнемдедік.

Осы мысалдардан біз мынандай қорытынды жасаймыз: бірқалыпсыз кодтың үнемділігі немесе тиімділігін жеке кодтық таңбалардың ұзындығы емес, олардың «орташа» ұзындығы $\underline{l}$ сипаттайды. Бірқалыпсыз кодтың орташа ұзындығы мынандай теңдеумен сипатталады:

$$\underline{l} = \sum_{i=1}^{N} l_i P(A_i)$$

Мұндағы l_i – A_i хабарламаның кодтық таңбасының ұзындығы, $P(A_i)$ – A_i хабарламаның ықтималдығы, N – хабарламаның жалпы саны. Ең үнемді кодтың орташа ұзындығы минимал шама болу керек. Енді 4 кестеде берілген бірқалыпсыз кодтың орташа ұзындығын табайық.

$$\underline{l} = 1 \text{ x } 0{,}5 + 2 \text{ x } 0{,}25 + 3 \text{ x } 2 \text{ x } 0{,}125 = 1{,}75$$

3 кестедегі қарастырылған бірқалыпты кодтың орташа ұзындығы $\underline{l} = 2$

«Артық» ақпарат ұғымы және оның кодтау амалындағы орны. Біз жоғарыда хабарға енетін белгілер тең емес ықтималдықта болса, онда оның ақпараттық энтропиясы азаятынын айтып кеткенбіз. Шеннон салыстырмалы ақпаратты енгізді. Ол H_1/H_m шамаға тең. Мұндағы H_1 хабар арқылы жеткен информация мөлшері, ал H_m хабарды құрайтын барлық белгілер тең ықтималды болған жағдайда хабармен жететін ақпарат мөлшері. Егер белгілер саны n болса $H_m = log_2 n$ болады. Ал салыстырмалы информацияны бір санына толықтыратын R шаманы «артық» ақпарат деп атады.

$$R=1 - \frac{H_1}{H_m}$$

Кейбір тілдердегі артық ақпарат қазіргі уақытта анықталып отыр: француз тілінде 55% артық ақпарат бар, ал орыс тілінде артық ақпарат 50%-ті құрайды. Тілдегі артық ақпарат канал арқылы тасымалданатын хабардың кодының ұзын болуына әкеліп соғады, дегенмен тілдің бұл қасиетінің арқасында хабар беру кезіндегі қателіктер де аз болады. Жюль Верннің «Капитант Гранттың балалары» атты романының кейіпкерлерінің бастан кешкен қилы оқиғаларының сәтті аяқталуы да тілдегі артық ақпараттың бар болуына да байланысты [7]. Қазақ тіліндегі артық ақпаратты анықтау барысында жүргізген кешенді зерттеуіміздің нәтижесінде тіліміздегі артық ақпарат 60%-ті құрайтынын анықтадық. Төмендегі 10% граматикалық белгілері алынып тасталған мәтін берілген. Сонда да осы хабарды оқып шығуға және түсінуге болады.

Мұғалім: Асқар, айтшы, егер бір қосу бір екі, ал екі қосу екі төрт, ал төрт қос төрт қанша болады ?

Асқар: Бұл әділетсіздік, мұғалім. Басқа оқушыларға оңай сұрық қоясыз, ал маған тек қиын сұрақтар ғана қалады.

Әр түрлі жанрларды қамтитын 50 мәтін жасалды. 40%, 50%, 55% және 60% белгілері алып тасталған және 300-500 белгілерден тұратын тапсырмаларды мектептің 11 сынып оқушылары мен жоғары оқу орны студенттеріне бердік. Алғыр оқушылар мен талантты студенттердің аз тобы 60 пайыз мөлшердегі әріптері жетпейтін тапсырмаларды орындап шықты.

Шындығында, әр түрлі қателікке толы мәтінді қиналмай-ақ оқып және мағынасын түсініп алатынымыз да тілдегі артық ақпараттың болуына тікілей байланысты.

Қазіргі кезде ғалымдар үнемділік мақсатында тілдегі артық ақпаратты түгел алып тастамай белгілі бір шекке дейін азайтып, хабар тасымалдаудың тиімді жағдайын жуық болса да тауып отыр [8]. Қазіргі уақытта хабарда кетіп қалған қателерді тауып алатын және оны түзететін арнайы кодтар негізінде жұмыс істейтін автоматтандырылған жүйелер пайда болды. Олардың жұмыс істеу принципі тілдердегі артық ақпарат факторына негізделген.

Қорытынды

Әлемдік қоғамдастық ақпараттық кезеңге аяқ басып отырған жағдайда ақпараттың қоғамдық қарым-қатынастағы, экономикадағы және ғылымдағы рөлі және орны ерекше болып отыр. «Кім ақпараттың иесі болса, ол барлық дүниенің иесі» түрінде айтылған қанатты сөздің мағынасымен қазір ешкім дауласа алмайды.

Математика және компьютерлік ғылымдардың үлкен салалаларының бірі ақпарат теориясының негізін құрайтын орташа ақпаратты өлшеуге арналған теңдеудің (Шеннон теңдеуі) хабарламаларды тығыздау мақсатында кодтау процедурасын зерттей отырып төмендегідей нәтижелер алдық:

- Кодтау технологиясының ақпараттарды тасымалдаудағы рөлі мен орны анықталды;
- Қазақ тіліндегі әріптердің мәтінде пайда болу жиілігі (ықтималдығын) есептелді;
- Мәтінді құрайтын элементтерді үнемді түрде кодтауда және байланыс каналдарымен тиімді тасымалдау мәселесін шешуде орташа ақпаратты өлшеуге арналған теңдеу қолданылды;
- Қазақ тіліндегі артық ақпаратты анықтауға арналған зерттеу жасалды.

Алынған теориялық және практикалық нәтижелерді математикада, информатикада, қазақша жазылған мәтіндерді кодтау үдерісінде пайдалануға болады. Алдағы уақытта кодтау теориясынын басқа салаларын, атап айтқанда қазақша хабарламаларды құпияландыру және хабарламалардағы кездейсоқ қателерді табу және түзету мәселелерін зерттеуді мақсат қойып отырмыз.

Пайдаланылған әдебиеттер тізімі:

1 *Шеннон К. Работы по теории информации и кибернетике. М.: Изд-во иностранной литературы, 1963. 830 с.*

2 *Вернер М. Основы кодирования. М.: Техносфера, 2004. 288 с.*

3 *Хэмминг Р.В. Теория кодирования и теория информации. М.: Радио и связь, 1983. 176 с.*

4 *Мукушев Б.А.,Турдина А.Б., Мукушев С.Б. Компьютерный метод определения относительных частот букв казахского, русского и английского языков (программа для ЭВМ). Авторское свидетельство. Астана: Министерство юстиции Республики Казахстан. №874. 21 июнь 2012 г.*

5 *Жанабаев З.Ж., Мукушев Б.А., Турдина А.Б., Мукушев С.Б. Использование информационной энтропии при контроле учебной деятельности обучающегося // Информатика и образование. - 2008. - №10.- С.120-123.*

6 *Mukushev B.A., Zheldybaeva B.S., Musatayeva I.S., Mukushev B.A., Kariev K.U., Turdina A.B. Formation of the scientific worldview in schoolchildren based on the inclusion of synergetic ideas in the content of education // Integratsiyaobrazovaniya Integration of education. 2018. T.22, No. 4. Pp. 632-646.) DOI: 10.15507 / 1991-9468.093.022.201804.632-647.*

7 *Урок-исследование «Поиск информации по различным информационным источникам» https://урок.рф/library/urokissledovanie_poisk_informatcii_po_razlichnim_inf_201450.html*

8 *Шавенько Н.К. Основы теории кодирования и сжатия сообщений: учебно-методическое пособие. М.: МИИГАиК, 2020. 87 с.*

References:

1 *Shannon K. (1963) Raboty po teorii informacii i kibernetike [Works on information theory and cybernetics. M.: Publishing House of Foreign Literature]. 830. (In Russian)*

2 *Werner M. (2004) Osnovy kodirovaniya [Fundamentals of coding]. M.: Technosphere. 288. (In Russian)*

3 *Hamming R.V. (1983) Teoriya kodirovaniya i teoriya informacii [Coding theory and information theory]. M.: Radio and Communications. 176. (In Russian)*

4 *Mukushev B.A., Turdina A.B., Mukushev S.B. (2012) Komp'yuternyj metod opredeleniya otnositel'nyh chastot bukv kazahskogo, russkogo i anglijskogo yazykov (programma dlya EVM) [Computer method for determining the relative frequencies of letters of Kazakh, Russian and English languages (computer program)]. Copyright certificate. Astana: Ministry of Justice of the Republic of Kazakhstan. No.874. 2012 (In Russian)*

5 *Zhanabaev Z.Zh., Mukushev B.A., Mukushev S.B., Turdina A.B. (2008) Ispol'zovanie informacionnoj jentropii pri kontrole uchebnoj dejatel'nosti obuchajushhegosja [The use of information entropy in the control of educational activity of a student]. Informatika i obrazovanie. №10, 120-124. (In Russian)*

6 *Mukushev B.A., Zheldybaeva B.S., Musatayeva I.S., Mukushev B.A., Kariev K.U., Turdina A.B. (2018) Formation of the scientific worldview in schoolchildren based on the inclusion of synergetic ideas in the content of education. Integratsiyaobrazovaniya, Integration of education. T.22, No. 4. 632-646. DOI: 10.15507/1991-9468.093.022.201804.632-647.*

7 *Urok-issledovanie «Poisk informacii po razlichnym informacionnym istochnikam» [Lesson-research "Search for information on various information sources"]. https://урок .рф /library/urokissledovanie_poisk_ informatcii_po_razlichnim_inf_201450.html (In Russian)*

8 *Shavenko N.K. (2020) Osnovy teorii kodirovaniya i szhatiya soobshchenij [Fundamentals of the theory of encoding and compression of messages: uchebno-metodicheskoe posobie] an educational and methodological manual. M.: MIIGAIK. 87. (In Russian)*