МАТЕМАТИЧНЕ ТА КОМП'ЮТЕРНЕ МОДЕЛЮВАННЯ

МАТЕМАТИЧЕСКОЕ И КОМПЬЮТЕРНОЕ МОДЕЛИРОВАНИЕ

MATHEMATICAL AND COMPUTER MODELLING

УДК 681.391

О. О. Архипова, В. М. Журавльов

ЧАСТОТНИЙ АНАЛІЗ ВИКОРИСТАННЯ БУКВ УКРАЇНСЬКОЇ МОВИ

В статті розглянуті невирішені задачі в галузі дослідження якості каналів мовної комунікації. Проведено частотний аналіз використання букв української мови для текстів загальним обсягом біля 580 тисяч знаків художнього, публіцистичного та технічного спрямування. Вперше отримано гістограму частот використання букв алфавіту для сучасної української мови.

ВСТУП. ПОСТАНОВКА ЗАДАЧІ

Якість передачі мовлення – це одна із головних характеристик каналу мовного зв'язку. Згідно [1], головними критеріями якості каналів мовної комунікації є:

1) розбірливість (зрозумілість, ясність);
2) гучність (голосність);
3) природність (натуральність).

Розбірливість, безперечно, можна назвати головним параметром, оскільки вона відображає виконання системою прийому-передачі мови свого головного призначення – забезпечення того, щоб слухач правильно зрозумів зміст переданого.

Гучність – визначає бажаний рівень прийнятих сигналів, який для оптимальних умов має бути таким, щоб не викликати стомленості та перенапруження слухового апарата. Цей параметр не є самодостатнім і використовується разом із першим, а в умовах використання спеціальної техніки, що регулює гучність, втрачає сенс.

Природність – оцінює здатність системи відтворювати не тільки зміст мови, що передається, але й її індивідуальні особливості, притаманні різним мовцям. Цей параметр не такий важливий як розбірливість. Виключенням є випадки спеціальних систем зв'язку, наприклад, систем, у яких потрібне визначення особи (ідентифікація) мовця за голосом, або для художнього відтворення мови та музики.

Усі відомі на цей час методи оцінки якості передачі мовлення можуть бути розділені на дві великі групи: *суб'єктивні експертні методи* і *об'єктивні методи* [1].

Артикуляційні випробування є суб'єктивним методом оцінки розбірливості – це найбільш прямий й очевидний, а іноді й єдиний, шлях дослідження якості каналу мовного зв'язку. Головними перевагами методу артикуляційних випробувань є його універсальність та відносна простота. Однак процедура організації суб'єктивних експертиз за оцінкою розбірливості мови – справа громіздка, тривала й досить дорога.

Серйозною й самостійною складною проблемою методу артикуляційних випробувань є створення спеціальних артикуляційних таблиць. Як показує практика вимірювань, тип таблиць, що використовуються, істотно впливає на результати вимірювань. Артикуляційні таблиці складаються за певними правилами [2]. Ці правила враховують лінгвістичні (мовні) і технічні вимоги до таблиць. Лінгвістичні вимоги полягають у тому, щоб таблиці достатньою мірою відображали фонетичну структуру мови. Технічні вимоги передбачають забезпечення максимальної економності під час виконання вимірів, мінімальної надмірності із максимально можливою однорідністю для того, щоб зменшити розкид результатів одиночних вимірів. Поєднання в таблицях лінгвістичних і технічних вимог можливе тільки у разі розумного компромісу, оскільки вони взаємно суперечливі.

Для української мови не складено артикуляційних таблиць, їх складання є актуальною науково-технічною задачею. Для забезпечення лінгвістичних вимог до артикуляційних таблиць для української мови необхідним є буквений та, у подальшому, фонемний частотні аналізи.

ЧАСТОТНИЙ АНАЛІЗ ВИКОРИСТАННЯ БУКВ УКРАЇНСЬКОЇ МОВИ

Кількість різних букв, як і фонем, у кожній мові обмежена. Важливими характеристиками мови є повторюваність букв (монограм), пар букв (біграм) і взагалі m-грам, сполучуваність букв одна з одною, чергування голосних і приголосних тощо. Примітно, що ці характеристики є досить стійкими [3].

Якщо $\vartheta(a_{i1}a_{i2}\ldots a_{im})$ – кількість появ m-грами $a_{i1}a_{i2}\ldots a_{im}$ у тексті T, а L – загальне число підрахованих m-грам, то при досить великих L частоти

$$\frac{\vartheta(a_{i1}a_{i2}\ldots a_{im})}{L} \qquad (1)$$

для даної m-грами мало відрізняються одна від одної.

Виходячи з цього, відносну частоту (1) вважають наближеною ймовірністю P $(a_{i1}a_{i2}\ldots a_{im})$ появи даної m-грами у випадково обраному місці тексту (за статистичним визначенням імовірності).

Частотний аналіз використання букв проведений для ряду європейських мов, його результати наведені у книзі [4]. Необхідно зазначити, що частота використання букв для французької, німецької, англійської іспанської та італійської мов різна. Деяка різниця значень частот у таблицях, які наводяться з різних джерел, пояснюється тим, що частоти істотно залежать не тільки від довжини тексту, але й від його характеру. Наприклад, у технічних текстах рідка буква Ф може стати досить частою у зв'язку із частим використанням таких слів, як функція, диференціал, дифузія, коефіцієнт і т. п.

Таблиця 1 – Частоти використання букв російської мови

–	О	Е, Ё	А	И	Т	Н	С
0.175	0,09	0,072	0,062	0,062	0,053	0,053	0,045
Р	В	Л	К	М	Д	П	У
0,04	0,038	0,035	0,028	0,026	0,025	0,023	0,021
Я	Ы	З	Ь, Ъ	Б	Г	Ч	Й
0,018	0,016	0,016	0,014	0,014	0,013	0,012	0,01
Х	Ж	Ю	Ш	Ц	Щ	Э	Ф
0,009	0,007	0,006	0,006	0,004	0,003	0,003	0,002

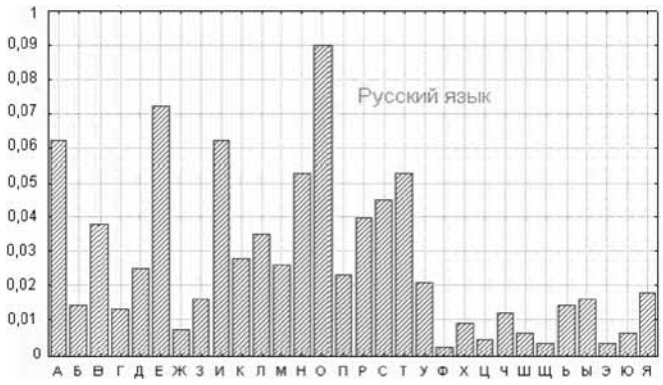

Рисунок 1 – Гістограма частот використання букв алфавіту російської мови

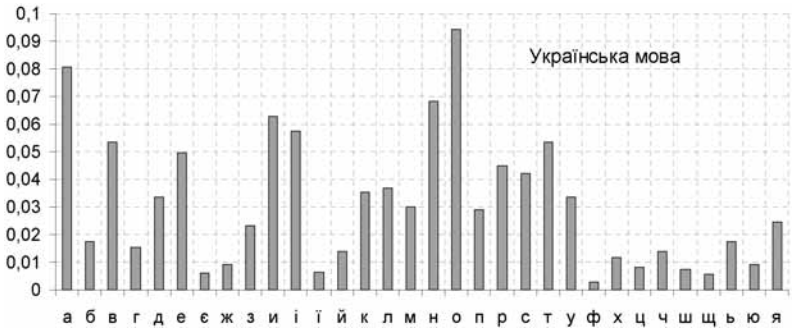

Рисунок 2 – Гістограма частот використання букв алфавіту української мови

Таблиця 2 – Осереднені частоти та дисперсії використання букв української мови

	Технічні тексти		Вірші		Гуманітарні твори	
	частота	дисперсія	частота	дисперсія	частота	дисперсія
А	0,0709	2,350E-05	0,0871	1,564E-04	0,0840	2,870E-05
Б	0,0136	1,092E-05	0,0211	6,685E-08	0,0183	4,918E-06
В	0,0533	2,066E-05	0,0468	1,573E-05	0,0604	9,286E-05
Г	0,0142	7,898E-06	0,0150	1,450E-05	0,0173	2,897E-06
Д	0,0350	3,305E-05	0,0332	9,669E-07	0,0332	3,025E-06
Е	0,0458	7,780E-05	0,0544	3,143E-07	0,0483	3,821E-05
Ж	0,0066	2,227E-06	0,0108	1,457E-06	0,0106	3,962E-06
З	0,0238	2,096E-05	0,0220	3,176E-06	0,0239	7,000E-06
И	0,0613	3,346E-05	0,0633	3,170E-05	0,0632	9,140E-07
Й	0,0114	8,932E-06	0,0162	4,442E-08	0,0139	7,501E-06
К	0,0358	1,231E-05	0,0330	1,674E-07	0,0373	4,538E-07
Л	0,0305	8,335E-06	0,0432	7,376E-06	0,0370	2,138E-05
М	0,0283	3,545E-05	0,0328	4,833E-08	0,0297	8,720E-06
Н	0,0836	1,155E-05	0,0562	6,087E-06	0,0645	3,203E-05
О	0,0950	2,055E-05	0,0920	3,702E-05	0,0956	9,878E-06
П	0,0304	3,369E-05	0,0264	1,494E-06	0,0303	2,258E-06
Р	0,0499	5,291E-05	0,0400	4,206E-06	0,0445	9,543E-06
С	0,0406	4,418E-05	0,0458	4,125E-06	0,0407	6,013E-06
Т	0,0546	8,941E-06	0,0576	6,375E-06	0,0483	5,564E-05
У	0,0321	2,617E-05	0,0333	1,083E-06	0,0353	5,434E-06
Ф	0,0061	1,096E-05	0,0004	1,239E-08	0,0020	1,871E-06
Х	0,0129	1,185E-05	0,0123	1,771E-06	0,0105	5,424E-07
Ц	0,0108	8,155E-06	0,0061	2,982E-06	0,0080	6,574E-06
Ч	0,0136	7,865E-06	0,0144	4,283E-09	0,0144	3,963E-06
Ш	0,0060	2,921E-06	0,0083	7,629E-07	0,0085	1,252E-06
Щ	0,0040	1,147E-06	0,0063	3,813E-06	0,0065	8,882E-07
І	0,0628	3,037E-05	0,0530	2,239E-04	0,0568	2,204E-05
Ї	0,0076	4,060E-06	0,0060	1,413E-06	0,0058	2,993E-06
Ь	0,0167	7,304E-06	0,0211	2,969E-06	0,0153	6,865E-06
Є	0,0080	5,295E-06	0,0062	1,243E-06	0,0042	3,730E-06
Ю	0,0082	4,378E-06	0,0116	4,322E-08	0,0081	1,980E-06
Я	0,0268	5,537E-05	0,0240	2,921E-06	0,0236	3,870E-06

Таблиця 3 – Ранжовані частоти використання букв української мови

О	0,0942	р	0,0448	я	0,0248	ж	0,0093
А	0,0807	с	0,0424	з	0,0232	ю	0,0093
Н	0,0681	л	0,0369	б	0,0177	ц	0,0083
И	0,0626	к	0,0354	ь	0,0177	ш	0,0076
І	0,0575	д	0,0338	г	0,0155	ї	0,0065
В	0,0535	у	0,0336	ч	0,0141	є	0,0061
Т	0,0535	м	0,0303	й	0,0138	щ	0,0056
Е	0,0495	п	0,0290	х	0,0119	ф	0,0028

Ще більші відхилення від норми в частоті вживання окремих букв спостерігаються в деяких художніх творах, особливо у віршах. Тому для надійного визначення середньої частоти букв бажано мати набір різних текстів, запозичених з різних джерел. Разом із тим, як правило, подібні відхилення незначні і, в першому наближенні, ними можна знехтувати.

Для російської мови частоти знаків алфавіту (у порядку зменшення), де ототожнено Е з Ё, Ь з Ъ, а також є знак пробілу (–) між словами, наведені в табл. 1 (див. [5]), або у вигляді наочної діаграми, приведеної на рис. 1.

Нами був проведений частотний аналіз повторюваності букв української мови за допомогою програми, написаної на мові програмування С++ та пакету Excel. У ході аналізу оброблено біля 580 тисяч знаків українських текстів. Серед них 260 тисяч – сучасні тексти технічного спрямування (роботи та статті по захисту інформації, математичній статистиці, диференційним рівнянням) 76 тисяч знаків – вірші (Т. Шевченко та сучасна поетеса Н. Доценко) та 253 тисяч – відома художня проза та публіцистика (твори Д. Лондона, Е. По, А. Гофмана, газета «Дзеркало тижня»). Осереднені частоти та дисперсії використання букв української мови наведені у табл. 2.

У табл. 3 містяться середні ймовірності появи букв українського алфавіту ранжовані у порядку спадання, а на рис. 2 – діаграма частот використання букв алфавіту української мови. У зв'язку із тим, що у багатьох текстах не розрізняються літери Г і Ґ, їх було ототожнено.

Порівнюючи гістограми частот використання букв російської, української та європейських мов [4], помітно різницю у їх розподілі, яка досягає десяти відсотків. Даний факт свідчить про ймовірне виникнення методичної похибки визначення розбірливості за словами та складами при використанні російськомовних артикуляційних таблиць для оцінки якості україномовних каналів мовної комунікації.

ВИСНОВКИ

1. Вперше сформульована науково-технічна задача створення артикуляційних таблиць для української мови.

2. Виконано необхідний проміжний етап цієї задачі, що забезпечує лінгвістичні вимоги до артикуляційних таблиць (відображення структури мови), – побудовано гістограму частот використання букв алфавіту сучасної української мови.

3. Отримана діаграма дозволяє скласти кілька артикуляційних таблиць з різним характером наборів слів (в залежності від тематичного спрямування) для проведення подальших артикуляційних випробувань. У майбутньому необхідно зробити фонемний частотний аналіз, також бажаним є мовний аналіз біграм і триграм.

ПЕРЕЛІК ПОСИЛАНЬ

1. *Покровский Н. Б.* Расчет и измерение разборчивости речи / Н. Б. Покровский. – М. : Связьиздат, 1962. – 392 с.
2. *Вемян Г. В.* Передача речи по сетям электросвязи / Г. В. Вемян. – М. : Радио и связь, 1985. – 272 с.
3. Обмен опытом [Електронний ресурс]. – Режим доступу: http://www.statsoft.ru/home/portal/exchange/textanalysis.htm. – Назва з екрана.
4. *Baudouin C.* Elements de cryptographie / C. Baudouin, Ed. A. Pedone. – Paris, 1939. – 214 p.
5. *Яглом А. М.* Вероятость и информация / А. М. Яглом, И. М. Яглом. – М. : Наука, 1973. – 374 с.

Надійшла 16.02.2009
Після доробки 27.04.2009

В статье рассмотрены нерешенные задачи в области исследования качества речевой коммуникации. Проведен частотный анализ использования букв украинского языка для текстов общим объемом 580 тысяч знаков художественного, публицистического и технического характера. Впервые получено гистограмму частот использования букв алфавита для современного украинского языка.

In this paper is considered unresolved tasks in the research field of the communications speech quality. The frequency analysis of the Ukrainian letters usage for total amount of texts about 580 thousand signs in art, publicistic and technical area is carried out. For the first time it is received the frequency histogram of letters usage of the Ukrainian alphabet for modern language.