УДК 004.93.1

В.М. Заяць*, М.М. Заяць
*Львівський державний інститут новітніх
технологій та управління імені В. Чорновола,
кафедра інформаційно-комп‘ютерних технологій та систем;
Національний університет “Львівська політехніка”,
кафедра інформаційних систем і мереж

МЕТОДИ ЗІСТАВЛЕННЯ СТАТИСТИЧНИХ ХАРАКТЕРИСТИК ПРИ ФОРМУВАННІ ВИБІРОК У ЛІНГВІСТИЦІ

Досліджуються загальні положення лінгвістичної статистики для мовознавця. Наведено конкретні приклади застосування методів зіставлення статистичних характеристик.

Ключові слова: лінгвістика, статистика, ймовірність.

The article deals with general problems of linguistic statistics from the point of view of general linguistics. The examples of methods application of statistical descriptions comparison are resulted.

Keywords: linguistic, statistics, probability.

Постановка проблеми

Сфера дії статистичних законів надзвичайно широка: усі складні системи підпорядковуються, передусім, законам статистичним. До таких систем належать живі організми та дослідження їх поведінки на основі імітаційних моделей, економіка та аналіз результатів діяльності великих колективів, методи інтерпретації лінгвістичних особливостей розвитку мови.

Важко знайти лінгвістичну роботу, в якій би не потрібні були елементарні підрахунки. Кількісні методи у лінгвістиці допомагають правильно організувати лінгвістичні спостереження, забезпечити надійність, точність, достовірність висновків у науці про мову. Ці методи ввійшли до мовознавчої практики, а лінгвістична статистика як наука існує вже 30 років.

У зв’язку зі сказаним проблема оптимального формування статистичних вибірок під час опрацювання великих масивів даних є актуальною.

Аналіз основних результатів

Активне використання математичних методів у вивченні мови почалося в середині XX ст. Стимулом для цього стали перспективи машинного перекладу. У процесі обробки текстів для їх уведення в машину було одержано різноманітні кількісні оцінки окремих фактів мови, які згодом виявилися корисними не тільки для створення математичних моделей мови, а й для лінгвістичної теорії. Оскільки мова — це ймовірнісна, а не жорстко детермінована система, то для її пізнання розроблено квантитативні методи, пов'язані з дослідженням частотних, ймовірнісних та інших нелогічних характеристик [5].

Розрізняють кількісні й статистичні методи. Кількісні методи зводяться до простого підрахунку частоти вживання мовних одиниць. Статистичні методи передбачають використання різних формул для виявлення правил розподілу мовних одиниць у мовленні, для виміру зв'язків між мовними елементами, для встановлення тенденцій у розвитку та функціонуванні мови та для встановлення залежності між якісними й кількісними характеристиками мови.

Лінгвістична статистика розглядається і як техніка обробки лінгвістичних даних, і як метод дослідження мови та мовлення, і як концепція, система ідей та уявлень про об'єкт лінгвістичної науки. І якщо техніку квантитативної лінгвістики використовують нині у багатьох дослідженнях, то методика реалізується тільки тоді, коли дослідник розуміє, що лише за допомогою кількісного

підходу можна одержати нові дані або перевірити набуті знання про лінгвістичний об'єкт, коли дослідник переконаний в імовірнісній природі лінгвістичного об'єкта і ставить перед собою завдання – описати його у кількісних показниках [6].

Можливість використання кількісних методів у мовознавстві ґрунтується на особливостях будови мови та мовлення. Мова – це система, яка складається з дискретних одиниць, що мають кількісні характеристики. Ці характеристики притаманні одиницям усіх рівнів. Разом з тим, мова має ймовірнісний характер: це код з імовірнісними обмеженнями. У загальному розумінні код – це засіб подання інформації у формі, здатній для передавання інформації каналами зв'язку. Будь-який код – це певна множина фізично різних знаків, кожен з яких може співвідноситися з тим чи іншим об'єктом з множини об'єктів, на які поширюється дія цього коду [7].

Мовлення є реалізацією системи мови, її елементів. Можна вказати на декілька факторів, що дають змогу застосовувати кількісні методи при дослідженні мовних та мовленнєвих даних:

1) дискретність одиниць;

2) масовість мовних одиниць (а у мовленні практично нескінченний ланцюжок);

3) повторюваність їх у висловлюваннях;

4) можливість вибору певного елемента з ряду однорідних. На мовлення впливають закони мови (закономірності будови одиниць мови, використання їх у мовленні), закони сполучуваності одиниць у мовленні, закони жанру, теми висловлювання, смаки автора, його психофізіологічний стан тощо. Дія цих факторів так переплітається, що інколи неможливо визначити результати їхнього впливу. Але якщо сукупність цих факторів порівняно постійна, то будова мовлення буде характеризуватися такими рисами, які можуть розкриватися кількісними методами.

Основним завданням статистичної лінгвістики є застосування кількісних методів для розкриття закономірностей функціонування одиниць мови у мовленні, а також у встановленні закономірностей будови тексту.

Перші роботи з розпізнавання образів стосувались теорії і практики побудови читальних автоматів (під образом розуміли знак, зображення, букву або цифру). Математичним апаратом для розв’язання задач розпізнавання з моменту їх виникнення була теорія статистичних розв’язків [13].

Сьогодні результати теорії статистичних розв’язків стали основою для побудови алгоритмів розпізнавання, які забезпечували зарахування об’єкта до його класу на підставі експериментальних апостеріорних даних – ознак, що характеризують об’єкт, та апріорних даних, що описують класи об’єктів. Пізніше математичний апарат розширився за рахунок використання методів алгебри логіки і деяких розділів прикладної математики, теорії інформації, математичного програмування і системотехніки [13–15].

Формулювання цілі статті

Метою статті є аналіз основних понять та інструментів математичної статистики, організації статистичних експериментів, основних методів та прийомів статистичного аналізу мовного матеріалу та формування статистичних вибірок у лінгвістиці.

Процедура вимірювання первинних ознак в процесі розпізнавання, встановлення пріоритету цих ознак та їх впливу на інтегральні характеристики досліджуваного об’єкта є дуже важливою під час створення критеріїв розпізнавання об‘єктів. З математичного погляду опис такої системи має забезпечувати мінімальну похибку розпізнавання та достовірну ідентифікацію об‘єкта розпізнавання за певними ознаками та критеріями прийняття рішення, що може бути реалізовано на основі теорії статистичних розв’язків.

Виклад основного матеріалу

Розпочинаючи статистичне дослідження матеріалу, мовознавець повинен чітко дати відповіді, принаймні, на три основні питання: що є об’єктом дослідження, з якою метою виконується аналіз, які методи застосовуються для розрахунків.

Відповідь на перше запитання передбачає визначення тих лінгвістичних одиниць, форм та категорій, які будуть статистично обстежуватися в експерименті.

Оскільки всі рівні мовної системи підпорядковуються дії статистичних законів, підрахункам можуть підлягати одиниці будь-якого рівня – фонеми або звуки, літери, сполуки звуків, фонем або літер, склади, морфеми, слова, словосполучення, синтаксичні конструкції тощо.

Відповідь на друге питання передбачає визначення мети і завдань, які стоять перед статистичним дослідженням.

Найповніша постановка задачі в статистиці – це визначення, істотні чи неістотні розходження між двома вибірками.

Відповідь на третє запитання передбачає розгляд та роз'яснення основних положень математичної статистики, що стосуються правил організації вибірки, визначення обсягу досліджуваного матеріалу для одержання статистично вірогідних результатів, обчислення статистичних характеристик, зіставлення статистичних параметрів різних сукупностей, статистичних методів визначення зв'язку досліджуваних одиниць у сукупності тощо.

Статистика грунтується на теорії випадкових подій. Будь-яка подія, настання і характер якої не можна точно визначити за попередніми подіями чи явищами, які можуть спричинити її, називається випадковою подією. Це не означає, що кожна з цих подій не залежить від якихось причин чи не є наслідком інших подій. Просто ми не можемо точно й однозначно визначити в кожному окремому випадку причину події, що відбулась, оскільки таких причин багато і дія кожної з них може змінюватися.

Розглянемо приклад з мовознавства. На вибір кожного слова, кожної синтаксичної структури і морфологічної форми для побудови висловлювання діють багато різних факторів, кожний з яких не є точно визначеним, може набувати різних значень. У результаті сукупної дії всіх цих факторів людина в одних випадках вибирає один спосіб висловити думку і побудувати висловлення, а в інших – ту саму думку виражає по-іншому. Однак у разі порівняно постійної сукупності факторів результати виявляються приблизно однаковими, тому ми можемо відрізнити стиль одного письменника від стилю іншого, встановити характерні особливості одного жанру порівняно з іншими тощо. Оскільки вибір кожної лінгвістичної одиниці ми не можемо передбачити однозначно, то він є випадковою подією.

Якщо ми візьмемо з тексту одного твору якогось автора певну кількість однакових відрізків, то частота однієї і тієї самої одиниці в цих відрізках не залишатиметься однаковою, оскільки рівнодійна впливу всієї сукупності умов побудови тексту весь час коливається і зумовлює непомітні (а іноді й помітні) простому спостереженню розбіжності у побудові висловлювання в різних відрізках тексту. В кожному окремому відрізку без підрахунків не можна точно передбачити частоту досліджуваного явища, тому вона теж є випадковою величиною.

Для того щоб можна було приблизно вказати інтервали, в яких може коливатися ця частота, необхідно статистично обстежити частину тексту, зробити відповідні підрахунки, визначити закон, якому підпорядковується розподіл одиниць у тексті. І лише після такої роботи можна з певною імовірністю передбачити і появу певної одиниці після іншої одиниці, і частоту, з якими ці одиниці можуть траплятись в подібних текстах.

Але для того, щоб висновки, які є результатом статистичного обстеження тексту, були вірогідними, необхідно, щоб ці тексти належали до однієї й тієї самої генеральної сукупності. Поняття генеральної сукупності – одне з важливих понять математичної статистики. Генеральною сукупністю називають однорідний масив деяких одиниць, які належить обстежити. Обсяг і характер генеральної сукупності залежать від постановки задачі дослідження.

Наприклад, якщо ми досліджуємо особливості стилю Івана Франка, то всі його твори становлять генеральну сукупність. Якщо дослідженню підлягає, скажімо, мова українських газет XX ст., то генеральну сукупність становлять абсолютно всі газетні тексти, видані в Україні за цей період. У разі дослідження структури слова в українській чи іншій мові генеральною сукупністю буде словниковий склад цієї мови.

Як бачимо, обсяг генеральної сукупності може значно змінюватися залежно від мети та завдань дослідження. В одних випадках генеральна сукупність порівняно невелика за обсягом, і вона вся підлягатиме статистичному дослідженню (наприклад, твори одного автора або словник), в

інших – межі її важко визначити точно, а значить, і суцільне її обстеження неможливе. Під час розв'язання задач, пов'язаних з дослідженням таких величин генеральних сукупностей, необхідно відібрати якусь кількість матеріалу з цієї генеральної сукупності, статистично його обстежити і на підставі одержаних результатів робити висновки про всю генеральну сукупність.

Найголовніша вимога до вибірки – вимога її репрезентативності.

Передбачається, що генеральна сукупність – це однорідний відносно досліджуваних одиниць і категорій масив матеріалу. Тому вимога однорідності висувається і при організації вибірки, і при визначенні генеральної сукупності.

У який би спосіб ми не організовували вибірку, незмінним залишається одне: ми виберемо з генеральної сукупності деяку кількість підвибірок – уривків тексту, сторінок, рядків тощо.

Оскільки статистика основана на так званому законі великих чисел, за яким результати статистичних підрахунків тим точніші й вірогідніші, чим більше матеріалу обстежено, то перша відповідь: чим більший обсяг матеріалу, тим точніші результати. В ідеалі кількість підвибірок повинна наближатися до нескінченності.

Але ніхто не досліджує нескінченну кількість матеріалу. Завдання полягає в тому щоб визначити той мінімум, який був би необхідним і достатнім для одержання надійних результатів, які б адекватно відображали дійсність.

Після того як вибірка організована, визначена величина підвибірки й вибірки, починається наступний етап статистичного дослідження — з'ясування основних статистичних характеристик, тобто тих кількісних показників, які характеризують поведінку досліджуваного явища в організованій нами вибірці, а саме визначення абсолютних частот, обчислення середньої частоти досліджуваної одиниці (математичного сподівання), знаходження статистичних оцінок середньої частоти: середнього квадратичного відхилення (дисперсії), міри коливання середньої частоти, коефіцієнта варіації, відносної похибки.

Абсолютною частотою називається число, яке вказує, скільки разів вжита певна одиниця у підвибірці (або у вибірці, якщо вона не ділиться на підвибірки). Значить, у разі правильної організації вибірки ми повинні одержати стільки абсолютних частот, скільки підвибірок містить наша вибірка.

Середня частота – це середньоарифметичне всіх частот. Значить, для знаходження середньої частоти треба додати всі абсолютні частоти і суму поділити на кількість підвибірок

$$\bar{x} = \frac{\sum x_i n_i}{\sum n_i},$$

де x_i – абсолютна частота i-ї варіанти підвибірок; n_i – кількість підвибірок, що мають абсолютну частоту x_i.

Абсолютні частоти завжди коливаються навколо середньої величини, відхиляючись від неї більшою чи меншою мірою. Тому для характеристики вживання тієї чи іншої одиниці в певному масиві недостатньо обчислити лише середню частоту. Треба ще показати, на яку величину можуть відхилятися від середньої абсолютні частоти. Таким показником є середньоквадратичне відхилення

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2 n_i}{\sum n_i}}.$$

Якщо абсолютні частоти досліджуваного явища підпорядковуються нормальному закону розподілу, то їх відхилення від середньої частоти визначено так: 68,3 % абсолютних частот містяться в межах $\bar{x} \pm \sigma$; 95,5 % абсолютних частот – у межах $\bar{x} \pm 2\sigma$; 99,7 % абсолютних частот – у межах $\bar{x} \pm 3\sigma$.

Як показують дослідження, справді, при правильній організації вибірки й достатньому обсягу досліджуваного матеріалу явища й одиниці, які часто з'являються в тексті, підпорядковуються нормальному закону розподілу. Одиниці, у яких менша частота, зазвичай підпорядковуються закону Пуассона. Цей закон є загальнішим порівняно з нормальним законом розподілу, нормальний закон – частковий випадок закону Пуассона. Тому статистичні показники застосовні і для закону

Пуассона. Отже, здебільшого при лінгвістичних дослідженнях можуть бути застосовані основні статистичні характеристики, серед яких середнє квадратичне відхилення, міра коливання середньої частоти, коефіцієнт варіації.

Коливання абсолютних частот може істотно вплинути на величину середньої частоти. Так, якщо при статистичному обстеженні певного масиву текстів ми внесемо у вибірку ті з них, які характеризуються вищими частотами (а це може трапитися при будь-якому способі організації вибірки), то і значення середньої частоти буде вищим, ніж тоді, коли для обстеження його було б узято з інших текстів.

Тому будь-яке статистичне дослідження, яке основане на вибірковому обстеженні генеральної сукупності, обов'язково передбачає оцінку можливих коливань середньої частоти. В результаті цієї оцінки дослідник одержує дані про можливі зміни середньої частоти при зміні вибірки у вибраній генеральній сукупності. Ці можливі зміни, що допускаються законами статистики, вважаються статистично неістотними, а тексти, що характеризуються середніми частотами, відмінність між якими статистично неістотна, можна об'єднати в одну вибірку і вважати статистично однорідними.

Обчислення міри коливання середньої частоти $\sigma_{\bar{x}}$ здійснюється за формулою:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}},$$

де $n = \sum n_i$

Ще однією дуже важливою характеристикою, яка використовується для зіставлення поведінки як різних одиниць в одній вибірці, так однієї і тієї самої одиниці у різних вибірках, є коефіцієнт варіації

$$v = \frac{\sigma}{\bar{x}}.$$

Коефіцієнт варіації v показує, яку частку середньої частоти становить середнє квадратичне відхилення, тобто ступінь стабільності вживання одиниці в тексті (чим нижчий коефіцієнт варіації, тим більша стабільність вживання).

Визначення статистичних характеристик певної вибірки – лише початок статистичного дослідження. Адже мета будь-якого статистичного дослідження полягає не в тому, щоб одержати ряд чисел, а в тому, щоб на підставі цих чисел зробити змістовні висновки, які дають змогу глибше проникнути у природу досліджуваного явища. Для мовознавця метою статистичного дослідження повинні слугувати лінгвістичні висновки про характер функціонування мовних одиниць і категорій у мовленні, про розбіжності й подібності у структурі різних мов. Інакше кажучи, аналіз однієї вибірки ще не дає матеріалу для потрібних висновків, для них необхідно зіставити характеристики кількох вибірок.

При дослідженні мовлення мета такого зіставлення – з'ясування того, чи залежить функціонування певних одиниць і категорій від різновиду мовлення, тобто від функціонального чи авторського стилю, жанру, тематики. Під час дослідження системи мови виникають питання, що стосуються однакових і різних рис у структурі мови в різні періоди її розвитку або у структурі різних мов у той самий період. Таке зіставлення потрібне для зіставної лексикографії, якщо зіставляються дані словників, чи для зіставної стилістики, якщо порівнюються дані з текстів певного стилю.

Обчислення міри коливань середньої частоти ($\sigma_{\bar{x}}$) дає змогу обчислити смугу коливання середньої відповідно до різних довірчих ймовірностей. Так, якщо смуга коливань становить $\bar{x} + \sigma_{\bar{x}}$, то вона охоплює 68,3 % середніх частот заданої генеральної сукупності, тобто обчислена з довірчою ймовірністю 68,3 %. Смуга $\bar{x} + 2\sigma_{\bar{x}}$ обчислена з довірчою ймовірністю 95,5 %, а смуга $\bar{x} + 3\sigma_{\bar{x}}$ – з довірчою ймовірністю 99,7 % (відоме правило трьох сигм).

Прийнявши бажану для нас довірчу ймовірність, ми можемо сформулювати нульову гіпотезу.

1. Якщо смуги $\bar{x} + \sigma_{\bar{x}}$ частот перетинаються між собою, то з ймовірністю 68,3 % можна стверджувати, що розходження зіставлюваних середніх частот неістотне, зумовлене статистично і ним можна нехтувати.

2. Якщо перетинаються між собою смуги частот $\bar{x} + 2\sigma_{\bar{x}}$, то з ймовірністю 95,5 % можна стверджувати, що розходження між ними зумовлене статистично й обидві вибірки належать до однієї генеральної сукупності.

3. Нарешті, якщо перетинаються смуги частот $\bar{x} + 3\sigma_{\bar{x}}$, то нульова гіпотеза приймається з довірчою ймовірністю 99,7 %.

Приклад 1. Дослідження частоти займенника ВІН у трьох вибірках дало такі результати: $\bar{x}_1 = 6{,}48$, $\sigma_{\bar{x}_1} = 1{,}159 \approx 1{,}16$; $\bar{x}_2 = 2{,}41$, $\sigma_{\bar{x}_2} = 0{,}8$; $\bar{x}_3 = 4{,}27$, $\sigma_{\bar{x}_3} = 1{,}049 \approx 1{,}05$. Визначити, чи істотне розходження між частотами займенника ВІН у трьох вибірках.

Перевіряємо спочатку першу гіпотезу і робимо висновок з довірчою ймовірністю 68,3 %.

Для цього визначаємо смугу коливань частоти для кожної вибірки і будуємо діаграму коливання смуг середньої частоти.

Для $\bar{x}_1$ – від 5,32 до 7,64.

Для $\bar{x}_2$ – від 1,61 до 3,21.

Для $\bar{x}_3$ – від 3,22 до 5,32.

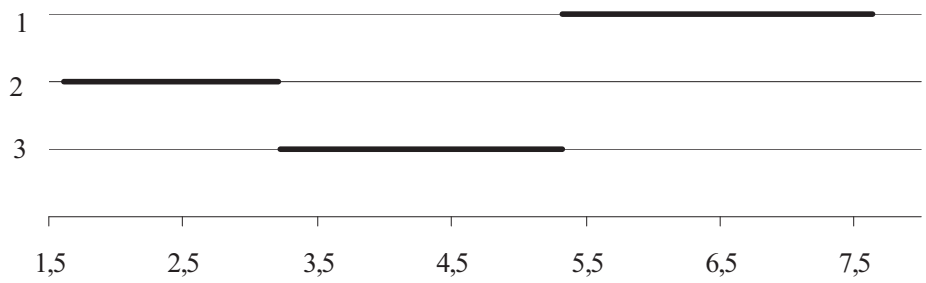

Рис. 1. Смуги коливань $\bar{x}$ при 68,3 % рівні довірчої ймовірності

Як показують діаграма і числові інтервали, жодна смуга частот не перетинається з іншими. Тому з ймовірністю 68,3 % можна стверджувати, що всі три вибірки істотно розходяться між собою за середніми частотами займенника ВІН, тобто нульова гіпотеза повинна бути відкинута.

Однак така низька довірча ймовірність твердження майже ніколи не приймається у статистичних дослідженнях: адже фактично третину можливих середніх частот вона не охоплює. Зазвичай у статистиці користуються 95 % довірчою ймовірністю.

Побудуємо інтервали середньої для цієї довірчої ймовірності: $\bar{x} + 2\sigma_{\bar{x}}$

Для $\bar{x}_1$ – від 4,16 до 8,8.

Для $\bar{x}_2$ – від 0,81 до 4,01.

Для $\bar{x}_3$ – від 2,17 до 6,37.

Як бачимо, тепер на смугах коливання середньої з'явилися частоти, спільні для $\bar{x}_1$ і $\bar{x}_3$ і для $\bar{x}_2$ і $\bar{x}_3$. Отже, середня частота в третій вибірці істотно не відрізняється ні від другої, ні від першої вибірки, тоді як між другою та першою вибіркою спостерігаємо істотну різницю. Це видно на діаграмі (рис. 2).

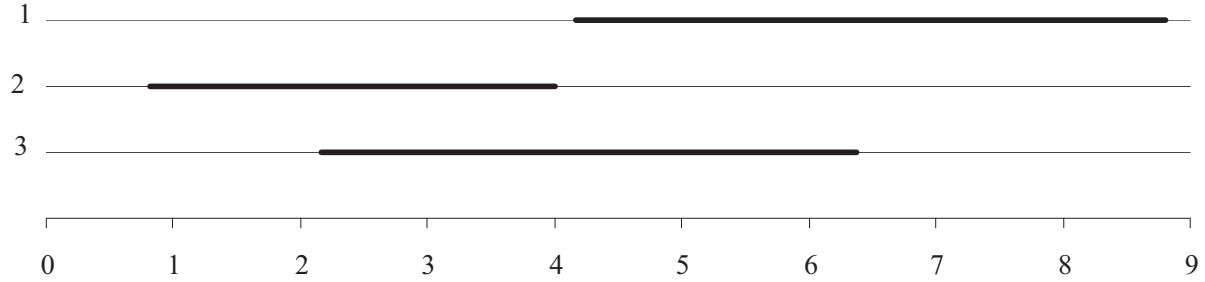

Рис. 2. Смуги коливань $\bar{x}$ при 95 % рівні довірчої ймовірності

Сформулюємо висновки з більшою точністю або з більшою (99 %) довірчою ймовірністю. Тоді смуги частот будуть такими:

для $\bar{x}_1$ – від 3 до 9,96;

для $\bar{x}_2$ – від 0,01 до 4,81;

для $\bar{x}_3$ – від 1,12 до 7,42.

Тепер між усіма смугами частот є ділянки перетину, тому з ймовірністю 99 % можна стверджувати, що всі три вибірки належать до однієї генеральної сукупності і що нульова гіпотеза може бути прийнята.

Будь-яка сукупність є однорідним масивом. Статистично однорідним певний масив текстів (або лінгвістичних одиниць) вважатиметься у тому випадку, коли досліджувані одиниці в усьому масиві мають статистичні характеристики, які істотно не відрізняються одна від одної. Так, якщо середня частота певного явища в одній вибірці істотно не відрізняється від його частоти в інших вибірках, то ці вибірки статистично однорідні відносно цього явища. Додавання слів "відносно цього явища" не випадкове, оскільки вибірки можуть бути статистично однорідними відносно одних явищ і різнорідними відносно інших. Тому визначення статистичної однорідності вибірок треба робити для кожного явища окремо. І не виключено, що групування одних і тих самих текстів буде різним для різних явищ або одиниць.

Для визначення статистичної однорідності треба мати ряди абсолютних частот для кожної вибірки. Будується таблиця, в якій стільки рядків, скільки вибірок зіставляється (кожний ряд визначено буквою М) і стільки стовпчиків, скільки підвибірок міститься у кожній вибірці (стовпчики позначені буквою К). Необхідно, щоб кількість підвибірок у зіставлюваних вибірках була однаковою.

Нехай маємо абсолютні частоти фонеми [а] у підвибірках по 1000 вживань кожна у трьох різних джерелах. Будуємо відповідну таблицю (табл. 1).

Таблиця 1

Обчислення χ^2

	K_1	K_2	K_3	K_4	K_5	∑M
M_1	98	103	99	78	94	472
M_2	85	82	75	91	71	404
M_3	102	88	71	97	85	443
∑K	285	273	245	266	250	1319

Розглянемо детальніше побудову таблиці. Перший рядок – це ряд абсолютних частот у першій вибірці й сума цих абсолютних частот (у стовпчику ∑M), другий – такі самі дані для другої вибірки. Третій – для третьої. Кількість таких рядків відповідає кількості порівнюваних вибірок. Останній рядок – це суми абсолютних частот у кожній підвибірці всіх порівнюваних вибірок. Число 1295 у правому нижньому кутку таблиці – це кількість вживань досліджуваного явища в усьому масиві (N).

Кожна абсолютна частота може бути позначена двома літерами з відповідними цифровими індексами. Так, число 98 позначається через K_1M_1, тобто перший стовпчик першого рядка, 99 – через K_3M_1, 80 – через K_2M_2 і т.д.

Критерій однорідності χ^2 обчислюється так:

$$\chi^2 = N\left(\sum \frac{(K_i M_j)^2}{\sum K_i \sum M_j} - 1\right),$$

де K_iM_j – абсолютна частота – j -го джерела i-ї підвибірки, i=1, ..., 5, а j=1, ..., 3.

Зробивши необхідні обчислення для нашої таблиці, отримаємо

$$\chi^2 = 9{,}6$$

Одержане кількість само по собі ще не свідчить про те, істотне розходження між частотами чи ні. Обчисливши кількість ступенів свободи:

$$f = (K-1)(M-1) = (5-1)(3-1) = 8,$$

у таблиці критичних значень χ^2 знаходимо значення: 15,5 і 20,1. Перше з них – це значення за довірчої ймовірності 95 %, а друге – при довірчій ймовірності 99 %. Якщо ми прийняли 95 % довірчу ймовірність і обчислений нами показник χ^2 менший від першого числа, то розходження між порівнюваними вибірками неістотне, і нульова гіпотеза приймається, якщо ж обчислений нами показник більший від табличного значення, то нульова гіпотеза відкидається. Наш показник χ^2 набагато менший від першого числа. Отже, з довірчою ймовірністю 95 % ми приймаємо нульову гіпотезу.

Як відомо, нульова гіпотеза приймається при $\chi^2 \le \chi^2_{05}$ і відкидається при $\chi^2 \ge \chi^2_{01}$. Це не значить, що нульова гіпотеза не може відкидатися, якщо показник, обчислений нами, міститься у проміжку, заданому обидвома числами таблиці. Просто у такому випадку ми маємо право стверджувати про істотне розходження з ймовірністю, не більшою за 95 %. У деяких випадках, що потребують особливо ретельної оцінки, така ймовірність може бути недостатньою. Якщо ж обчислене нами значення χ^2 перевищує χ^2_{01}, ми одержуємо повне право стверджувати про істотність розходження.

Описане тут обчислення статистичної однорідності має ту перевагу, що воно дає змогу зіставити будь-яку кількість вибірок. Однак така перевірка на однорідність дає змогу лише отримати відповіді “так” чи “ні” на питання щодо нульової гіпотези. Ані визначити, які саме вибірки розходяться, а які ні, ані обчислити ступені їх розходження цей метод не дає змоги.

Для відповіді на такі питання застосовується інша перевірка – перевірка за критерієм Стьюдента.

Для визначення істотності або неістотності розходження середньої частоти у двох вибірках за допомогою критерію Стьюдента достатньо знати абсолютні й середні частоти досліджуваної одиниці у цих вибірках. Тоді показник критерію Стьюдента (t) знаходимо за формулою:

$$t = \frac{|\bar{x} - \bar{y}|}{\sqrt{\frac{\sum (x_i - \bar{x})^2 n_i + \sum (y_i - \bar{y})^2 n_i}{n+m-2} \cdot \frac{n+m}{n \cdot m}}},$$

де $\bar{x}$ – середня частота досліджуваної одиниці у першій вибірці, а $\bar{y}$ – середня частота цього самого явища у другій вибірці; x_i – абсолютні частоти у першій вибірці; y_i – абсолютні частоти у другій вибірці; n – кількість підвибірок у першій вибірці; m – кількість підвибірок у другій вибірці.

Цю формулу можна значно спростити, якщо обчислювати у кожній вибірці стандартну похибку відхилення середньої $S_{\bar{x}} = \frac{\sigma}{\sqrt{n-1}}$:

$$t = \frac{|\bar{x} - \bar{y}|}{\sqrt{S^2_{\bar{x}} + S^2_{\bar{y}}}}.$$

Розглянемо приклад:

Приклад 1. Дослідження частоти іменника СЛОВО у двох прозових творах різних авторів наведено у табл. 2 та 3 (підвибірка – 1000 слововживань)

Таблиця 2

Частота іменника СЛОВО у першому творі

x_i	n_i	$x_i n_i$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 n_i$
0	18	0	-1,57954545	2,49496384	44,9093492
1	29	29	-0,57954545	0,33587293	9,74031508
2	22	44	0,420454545	0,17678202	3,88920455
3	13	39	1,420454545	2,01769112	26,2299845
4	4	16	2,420454545	5,85860021	23,4344008
5	1	5	3,420454545	11,6995093	11,6995093
6	1	6	4,420454545	19,5404184	19,5404184
∑	88	139			139,443182

Таблиця 3

Частота іменника СЛОВО у другому творі

x_i	n_i	$x_i n_i$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 n_i$
0	20	0	-1,47297297	2,16964938	43,3929876
1	22	22	-0,47297297	0,22370343	4,92147553
2	17	34	0,527027027	0,27775749	4,72187728
3	10	30	1,527027027	2,33181154	23,3181154
4	3	12	2,527027027	6,3858656	19,1575968
5	1	5	3,527027027	12,4399196	12,4399196
6	1	6	4,527027027	20,4939737	20,4939737
∑	74	109			128,445946

У першій вибірці $\bar{x} = 1{,}58$, у другій $\bar{y} = 1{,}47$

$$t = \frac{|1{,}58 - 1{,}47|}{\sqrt{\frac{139{,}44 + 128{,}45}{88 + 74 - 2} \cdot \frac{88 + 74}{88 \cdot 74}}} \approx 0{,}72 .$$

Щоб вирішити, чи про істотне розходження свідчить обчислений показник *t*, треба його оцінити за таблицею критичних значень *t*.

Оцінку цю виконують за допомогою визначення кількості ступенів свободи, яке у нашому випадку $f = 88 + 74 - 2 = 160$. Нульова гіпотеза відкидається, якщо обчислене значення *t* більше від табличного для заданого рівня значущості (довірчої ймовірності). В нашому прикладі 0,72 менше від найменшого числа у ряду. Це значить, що нульова гіпотеза приймається, тобто розходження у частоті іменника СЛОВО у досліджуваних творах неістотне, воно статистично допустиме і викликане звичайним коливанням обчислювальних частот.

Застосування критерію Стьюдента дає ті самі результати, що і зіставлення смуги коливання середньої частоти. Правда, оскільки у критерії Стьюдента застосовується $S_{\bar{x}}^2$, а не $\sigma_{\bar{x}}$, то при невеликих вибірках можуть бути розходження: зіставлення смуги коливання показує, що смуги не перетинаються, а критерій Стьюдента не показує істотного розходження.

Перевага критерію Стьюдента в тому, що за наявності істотних розходжень можна обчислити відносне розходження або ступінь розходження, відстань між зіставлюваними масивами за цією ознакою:

$$l = \frac{t - t_\alpha}{t},$$

де *t* – обчислене нами значення показника, а t_α – його критичне значення за заданої довірчої ймовірності.

Якщо за критерієм Стьюдента немає істотного розходження між зіставлюваними вибірками, то це свідчить про приналежність їх до однієї генеральної сукупності і дає нам право об‘єднати їх в одну вибірку.

Об’єднання вибірок потребує виконання нових підрахунків для обчислення середньої частоти і її статистичних характеристик.

У разі об’єднання вибірок має значення величина кожної з об’єднуваних вибірок: чим більша вибірка, тим більшу вагу матимуть її результати у новій вибірці, куди вона входить разом з якоюсь невеликою вибіркою. Тому при формуванні нової вибірки треба врахувати той вплив, який може чинити на неї кожна зі старих вибірок.

Цей підхід до формування вибірок застосований під час реалізації комп’ютерної системи розпізнавання зразків почерку користувачів комп’ютера [3]. Похибка розпізнавання при цьому не перевищує 5 %, якщо зареєстровано 200 користувачів.

Висновки

У роботі досліджено методи формування вибірок за допомогою зіставлення різних статистичних характеристик та перевірки ефективності вибірки за критерієм Стьюдента стосовно задач математичної лінгвістики. Ефективність методу підтверджено на конкретних прикладах з прикладної лінгвістики та під час реалізації системи ідентифікації користувачів комп‘ютера за їх рукомоторними реакціями.

Подальші дослідження пов’язані з розробленням ефективних алгоритмів автоматичного формування статистичних вибірок під час розв’язанні лінгвістичних задач.

1. Заяць В.М. Методи, алгоритми та програмні засоби для моделювання та аналізу динаміки складних об‘єктів та систем на основі дискретних моделей / Заяць В.М.– Львів: Новий світ – 2009. – 400 с. 2. Платонов А.В. Використання експертних ситуативних моделей у сфері державної безпеки / А.В. Платонов, І.В. Баклан, К.В. Крамер // Зб. праць міжнар. наукової конф. ISDMCI' 2008. Євпаторія. – 2008. – Т. 1. – С. 39–43. 3. Заяць В.М. Перспективні напрямки розвитку та застосування систем розпізнавання та ідентифікації об‘єктів і процесів на основі дискретних моделей коливних систем / В.М Заяць, М.М. Заяць // Вісник Нац. ун-ту ,,Львівська політехніка” "Інформаційні системи та мережі". – 2008. – № 610. – С.137–147. 4. Заяць В.М. Построение и анализ модели дискретной колебательной системы / В.М. Заяць // Кибернетика и системный анализ. – 2000. – С. 161–165. 5. Перебийніс В.І. Статистичні методи для лінгвістів / В.І. Перебийніс. – Нова книга. – 2002. – 170 с. 6. Заяць В.М. Аналіз динаміки та умов стійкості дискретних моделей коливних систем / В.М. Заяць // Вісник Нац. ун-ту ,,Львівська політехніка” "Інформаційні системи та мережі". – 2004. – № 519. – С.132–142. 7. Щербина Ю.М. Предмет математичної лінгвістики / Ю. Щербина // Вісник Нац. ун-ту ,,Львівська політехніка” "Інформаційні системи та мережі". – 2008. – № 631. – С.138–147. 8. Заяць В.М. Алгоритмічне та програмне забезпечення системи розпізнавання людини за її рукомоторними реакціями / В.М. Заяць, О.О. Уліцький // Вісник Держ. ун-ту ,,Львівська політехніка” "Комп’ютерна інженерія та інформаційні технології". – 2000. – № 392. – С.73–76. 9. Фукунага К. Введение в статистическую теорию распознавания / К. Фукунага. – М.: Наука, 1979. – 512 с. 10. Горелик А.Л. Методы распознавания / А.Л. Горелик, В.А. Скрипник. – М.: Высшая школа, 1989. – 232 с. 11. Дуда Р. Распознавание образов и анализ сцен / Р. Дуда, П. Харт. – М.: Мир, 1976. – 512 с. 12. Статистичні методи дослідження мовного матеріалу // У зб. «Методи структурного дослідження мови». – К.: Наукова думка, 1968. – С.120–163. 13. Заяць В.М. Математичний опис системи розпізнавання користувача комп’ютера / В.М Заяць, М.М. Заяць // Зб. "Фізико-математичне моделювання та інформаційні технології". – Львів. – 2005. – Вип. 1. – С. 146–152. 14. Заяць В.М. Підходи до побудови систем захисту інформації від несанкціонованого доступу / В.М Заяць, М.М. Заяць // Вісник Нац. ун-ту ,,Львівська політехніка” "Інформаційні системи та мережі". – 2008. – № 631. –С.138–147. 15. Zayats V. Statistic Results of Recognition Structural Method of Written Text by hand / V. Zayats, M. Zayats, D. Ivanov // Pros. of the Xth International Conf. "The experience of designing and application of CAD systems in microelectronics". – Lviv–Polyana. – 2009. – P. 494–496.