Наукові записки. Серія "Філологічні науки". Випуск 95(2). – К., 2011. – С. 538-542.
7. Ножов И. Морфологическая и синтаксическая обработка текста (модели и программы) [Електронний ресурс] / И. М. Ножов. – М., 2003. – 140 с. – Режим доступу: http://aot.ru/docs/Nozhov/msot.pdf.
8. Перебийніс В. Математична лінгвістика. Навчальний посібник / В. І. Перебийніс. – К.: Вид. центр КНЛУ, 2014. – 125 с.
9. Brooks A.Unsupervised Part-of-Speech Tagging: An Introduction [Електронний ресурс] / A. Brooks,
10. M. Stees. – Режим доступу: https://ou.monmouthcollege.edu/_resources/pdf/academics/mjur/2014/Unsupervised-Part-of-Speech-Tagging.pdf
11. Blevins J. Word-based morphology [Електронний ресурс] / J. P. Blevins // Journal of Linguistics. – Режим доступу: https://www.researchgate.net/publication/231781491_Word-based_morphology
12. Manning C. Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics? / C. D. Manning // Computational Linguistics and Intelligent Text Processing. CICLing 2011. Lecture Notes in Computer Science, vol 6608. – Springer, Berlin, Heidelberg. – P. 171-189.

Використання статистичних методів описової статистики у корпусній лінгвістиці

Зоя Кочуєва

к. т. н., професор, доцент кафедри інтелектуальних та комп'ютерних систем, Національний технічний університет «Харківський політехнічний інститут», Україна, E-mail: kochueva@kochuev.com

Валерій Дідусьов

магістр кафедри інтелектуальних та комп'ютерних систем, Національний технічний університет «Харківський політехнічний інститут», Україна, E-mail: valeradidusiov@gmail.com

The purpose of this study: the development of software to determine the criteria of Pearson Hi^2 on the basis of both scientific and literary texts. In the first part of the study it is conducted the analysis of contemporary corpus linguistics. Statistical studies and their views of corpus linguistics are also described. Held his own description of the body of fiction and non-fiction texts. In the second part of the paper it is focused on statistical methods in corpus linguistics, namely the descriptive statistics. The description of the operation of the developed software.

Ключові слова — комп'ютерні науки, лінгвістичні аспекти, корпус, критерії, генеральна сукупність, вибірка.

I. Вступ

Наукові дослідження в рамках різних лінгвістичних напрямків мають об'єктом свого дослідження текст або зібрання текстів і мають на увазі спочатку підбір матеріалу, а потім аналіз і обробку великих текстових обсягів з метою виявлення деяких мовних закономірностей. Традиційні лінгвістичні методи аналізу тексту дозволяють виконати всі перераховані вище завдання, але їх невисока ефективність обумовлює все більш часте використання методів комп'ютерного аналізу тексту, який дозволяє скоротити роботу лінгвіста, при цьому значно збільшити обсяг оброблюваних даних, а також уникнути неточності і помилки в підрахунках. Таким чином, комп'ютерний аналіз тексту уможливлює встановлення мовних закономірностей, заснованих не так на теоретичних, як на емпіричних даних [4].

У межах досліджень із корпусної лінгвістики під «корпусом» розуміють «уніфікований, структурований і розмічений масив мовних даних в електронному вигляді, призначений для певних філологічних і гуманітарних досліджень». Перевага застосування корпусного аналізу тексту полягає в великій мірі в об'єктивності дослідження: такі функції як підбір, розмітка, аналіз текстів та виявлення відповідностей виконуються автоматично. Таким чином, завданням дослідника є не аналіз матеріалу, а обробка отриманих даних, виведення мовних закономірностей і підведення підсумків.

Розвиток корпусної лінгвістики і зростання уваги до статистичних методів обробки мовного матеріалу за останнє десятиліття привели до розробки цілого ряду методик, пов'язаних з обробкою паралельних або близьких текстів різними мовами. Актуальною задачей є застосування до таких текстів методів статистики та лінійного програмування, які радикально скорочують трудомісткість робіт. У рамках даного дослідження увага приділяється застосуванню методів описової статистики, а саме методу для знаходження значення критерію Хі^2 для підтвердження гіпотези щодо відношення тексту до відповідного корпусу.

II.Виклад основного матеріалу дослідження

У практичній і науковій діяльності часто для доведення справедливості того або іншого факту удаються до висловлювання гіпотез, які можуть бути перевірені на основі даних вибіркового спостереження. Проаналізувавши статистичні методи було вирішено визначити мету даного дослідження - це використання статистичних методів описової статистики та перевірка статистичних гіпотез у корпусній лінгвістиці.

Перевірка статистичних гіпотез має велике значення для практики. Зокрема, на ній ґрунтуються прийоми статистичного контролю якості продукції. Припустимо, що на підприємстві про якість продукції роблять висновки за результатами вибіркового контролю. Якщо вибіркова частка браку не перевищує заздалегідь встановленої (нормативної) величини, то партія продукції приймається. Однак висновок щодо відповідності якості продукції встановленим вимогам робиться на основі вибіркової перевірки і тому носить імовірнісний характер. Таким чином, судження про якість продукції не може розглядатися як категоричне. По суті, мова йде про припущення (гіпотезу), що частка, браку у всій генеральній сукупності дорівнює або менше нормативної величини. Ця гіпотеза і має бути перевірена на основі результатів вибіркового спостереження. Перевірка гіпотез проводиться за допомогою методів математичної статистики.

Гіпотеза в широкому розумінні - це деяке наукове припущення щодо властивостей явищ, що їх вивчають, яке потребує перевірки та доведення.

Статистичною гіпотезою називається припущення відносно параметрів або форми розподілу генеральної сукупності, яке перевіряється на основі даних вибіркового спостереження. Позначається гіпотеза літерою Н від латинського слова hypothesis. Із визначення статистичної гіпотези випливає, що вона може стосуватися або окремих параметрів розподілу, або законів розподілу [3].

Наведемо загальну схему (алгоритм) перевірки статистичної гіпотези. Ця перевірка, як зазначалося вище, може бути проведена з використанням параметричних і непараметричних критеріїв. Наведемо схему перевірки гіпотези, що передбачає знання закону розподілу генеральної сукупності, тобто для випадку застосування параметричних критеріїв.

Перевірка цієї статистичної гіпотези передбачає послідовне виконання таких етапів:

1. Оцінка вихідної інформації та описування статистичної моделі вибіркової сукупності.
2. Формулювання нульової і альтернативної гіпотез.
3. Встановлення рівня значущості, за допомогою якого контролюється помилка I роду.
4. Вибір найпотужнішого критерію для перевірки нульової гіпотези. Застосування найпотужнішого критерію дозволить контролювати ймовірність появи помилки II роду.
5. Обчислення за певним алгоритмом фактичного значення критерію.
6. Визначення критичної області та області згоди з нульовою гіпотезою, тобто встановлення табличного значення критерію.
7. Зіставлення фактичного і табличного значень критерію і формулювання висновків за результатами перевірки нульової гіпотези.

Прийоми перевірки статистичних гіпотез залежать від характеру формування вибіркових сукупностей.

Формування вибіркових сукупностей зумовлює різні прийоми оцінки вірогідності між середніми двох малих вибірок. Якщо вибірки незалежні, то статистичній оцінці підлягає різниця середніх, якщо залежні - середня різниця [1].

Такі гіпотези, як і гіпотези відносно параметрів розподілу, перевіряють за допомогою спеціальних критеріїв згоди.

Критерієм згоди називають критерій перевірки гіпотези щодо передбачуваного закону невідомого розподілу в генеральній сукупності.

Є ряд критеріїв згоди: К. Пірсона, О.М. Колмогорова, М.В. Смирнова, Б.С. Ястремського та ін. Ці критерії дозволяють встановити, узгоджуються чи не узгоджуються досліджувані розподіли з теоретичними розподілами, а також те, наскільки істотними є розбіжності між цими розподілами [2].

Для того щоб проводити деякі статистичні дослідження та застосовувати статистичні методики в корпусній лінгвістиці, необхідно мати репрезентативний корпус, що відповідає поставленому завданню. У рамках дослідження використовувався корпус «Наукова та художня література», з метою глибшого дослідження за допомогою даного корпусу текстових файлів та методу перевірки статистичних гіпотез у корпусній лінгвістиці. Використовуваний корпус являє собою вибірку, тобто містить лише деяку частину необхідного матеріалу на відміну від генеральної сукупності. Оскільки для даного дослідження використовувати усю генеральну сукупність не є можливим, у репрезентативній вибірці є усі елементи генеральної сукупності, а об'єкти, що часто проявляються в генеральній сукупності, частіше проявляються і в ній. Метою створення цього корпусу полягає в створенні максимально інформативної вибірки на цю тематику для подальшого статистичного аналізу.

У корпус було введено тексти двох жанрів літератури: художнього та наукового. Як вказано на Рис 1, основою структури корпусу служать 2 теки, що відповідають за напрям.

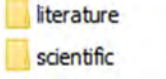

Рис. 1. Структура корпусу

Кожнен напрям розподілено на жанри. Це зображено на Рис. 2.

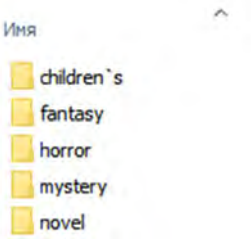

Рис. 2. Структура теки literature

У кожній з тек зберігається по 3 файли у форматі .txt, які відповідають за жанр літератури. Рис. 3.

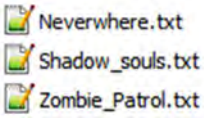

Рис. 3 Структура теки horror

У назві файлів вказана назва тексту з якого було взято частину. Кожен файл містить в середньому від 1500 до 1800 слів. Текстова інформація вибиралася по усій структурі текстової статті. Тобто кожен текстовий файл містить уривки зі вступної частини, кульмінації та кінцівки.

Загалом, загальний обсяг корпусу 40030 слів. Такий розмір задовольняє мінімальній вимозі до обсягу корпусу.

Основний принцип побудови корпусу полягає в тому, щоб усі елементи генеральної сукупності мали рівні шанси потрапити у вибірку. Але як би ретельно не дотримувалися цього принципу, випадкові помилки все ж матимуть місце у вибірці.

Для рішення даної задачі було створено програмний продукт призначений для перевірки статистичних гіпотез. У дослідженні у якості нульової гіпотези було обрано гіпотезу, що обраний текст відповідає корпусу художньої або наукової літератури.

На Рис. 4 зображено головне вікно програми

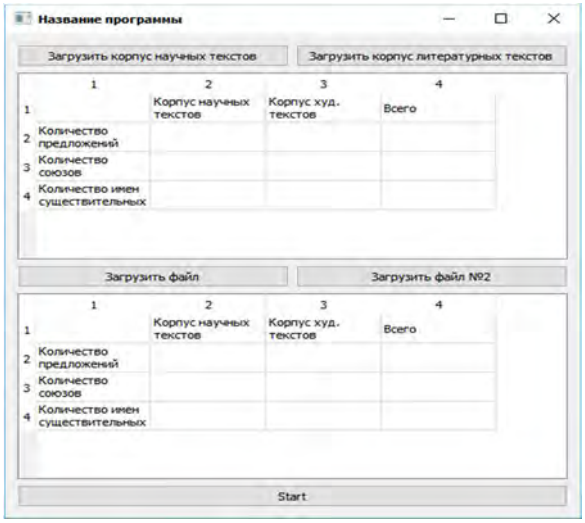

Рис. 4 Головне вікно програми

На Рис. 5 зображено результат роботи програмного продукту. Значення критерію, яке ми отримали у результаті роботи ми повинні порівняти із критичним значенням при різних рівнях значимості. Якщо критичне значення відповідного рівня значущості більше за отриманий результат то ми можемо зробити висновок, що гіпотеза підтверджується.

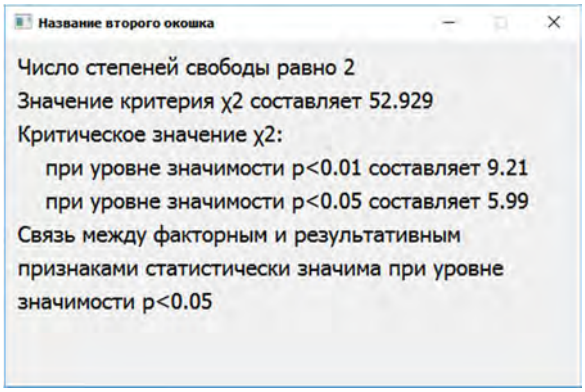

Рис. 5. Результат роботи програмного продукту

Висновок

Досліджуваним матеріалом був корпус, що містить збірку наукових та художніх текстів. Перевірка статистичних гіпотез здійснювалась відповідно до визначеного критерію Xi^2 Пірсона.

Знаючи такі величини, як кількість речень, кількість іменників, кількість сполучників, можна завдяки розробленому програмному продукту перевірити статистичні гіпотези. Проведений аналіз дозволив визначити нульову гіпотезу. У результаті роботи програми користувач може ознайомитись з результатами обробки. Основний показник — це значення критерію Xi^2. Його зіставляють з критичним значенням.

Література

1. Корпусна лінгвістика / В. А. Широков, О. В. Бугаков, Т. О. Грязнухіна та ін. – К.: Довіра, 2005. – 471 с.

2. Ивановский Р. Теория вероятностей и математическая статистика. Основы, прикладные аспекты с примерами и задачами в среде Mathcad. — 528 с.

3. Основые задачи и методы математической статистики: [Електронний ресурс]. – Режим доступу:http://uchebnikionline.com/statistika/matematicheskaya_statistika_htm.

4. Корпусная лингвистика: [Електронний ресурс].– http://www.myfilology.ru/177/korpusnaya-lingvistika-kak-razdel-yazykoznaniya/

5. Корпусная лингвистика: [Електронний ресурс]. – http://corpora.iling.spb.ru/theory.htm

6. Критерій Xi2 Пірсона: [Електронний ресурс]. – http://medstatistic.ru/theory/hi_kvadrat.html