Лекція 2

ЕЛЕМЕНТАРНІ ТА СКЛАДНІ ЛІНГВІСТИЧНІ ПОДІЇ. ОПЕРАЦІЇ НАД ЛІНГВІСТИЧНИМИ ПОДІЯМИ

1 ЛІНГВІСТИЧНА ПОДІЯ

1.1 Спостереження, випробування та подія в індуктивних дослідженнях мови

Основою всіх індуктивних досліджень у мовознавстві є спостереження за поведінкою і ознаками **лінгвістичних** об'єктів, що вивчаються. Це спостереження може здійснюватись також через експеримент або кількісне вимірювання. Здіснення кожного такого спостереження (досліду або вимірювання) називається випробуванням. Сукупність умов, за яких здійснюється дане випробування, називають *комплексом умов*, і позначають через σ.

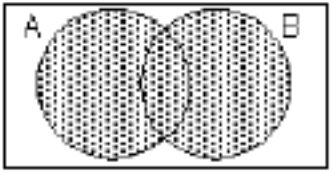
Результатом лінгвістичного випробування є *лінгвістична подія*.

Кожна подія, яка може відбутись, а може і не відбутись, називається *випадковою подією*. Якщо результат лінгвістичного випробування повністю вичерпується деякою однією (і тільки однією) подією, то маємо справу з *елементарною* випадковою подією. Подія, яка складається з декількох елементарних подій, означується як *складна* випадкова подія.

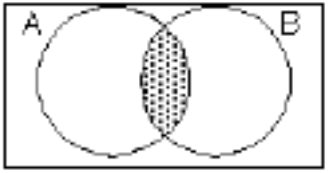
Проведемо дослід (випробування), який полягає у вгадуванні букви при такому комплексі умов σ_1: перед буквою, що вгадується, міститься ланцюжок *котр*, текст український без помилок. Це випробування може дати події A_1, B_1, C_1, D_1, E_1, які полягають, відповідно, в появі таких букв: *а* (*котра*), *е* (*котре*), *и* (*котрий*), *о* (*котрому*, *котрого*), *і* (*котрі*). Появи букв *а*, *е*, *и*, *о*, *і* після ланцюжка *котр* є елементарними випадковими подіями, появи після того ж ланцюжка діграм *ий*, *ом*, *ог* потрібно розглядати як складні випадкові події.

1.2 Операції над лінгвістичними подіями

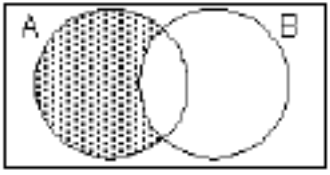
1. Складна подія, яка полягає у здійсненні хоча б однієї з подій A, B, називається *сумою* цих подій; позначається через $A+B$ (читається „A або B"). Поява букви *а* (подія A) або букви *е* (подія B) після ланцюжка *котр* є сумою $A+B$.



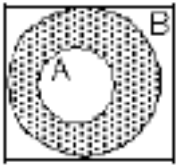
2. Складна подія, яка полягає у одночасному здійсненні подій A та B, називається їхнім *добутком*; позначається через AB (читається „A і B").



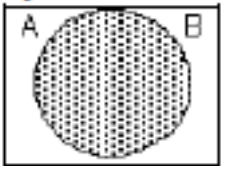
3. Складна подія, яка полягає у тому, що подія A відбувається а подія B не відбувається, називається різницею подій A та B; позначається через $A-B$.



4. Якщо подія A, яка відбувається при реалізації комплексу умов σ, викликає щоразу появу події B, то кажуть, що A є *частинним випадком* B, і записують $A \subset B$ (або $B \supset A$).



5. Якщо подія A при комплексі умов σ викликає появу події B і, навпаки, при цьому ж комплексі умов B викликає A, то події A та B називають рівносильними і записують $A = B$.



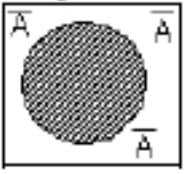
6. Якщо деяка подія при даному комплексі умов σ повинна обов'язково відбутись, то така подія називається *достовірною*. Подія, яка при комплексі умов σ відбутись не може, називається *неможливою*. Оскільки всі достовірні події рівносильні, їх прийнято позначати через U, неможливі події через ці ж міркування позначаються буквою V; $V = \overline{U}$.

7. Дві події називаються *несумісними*, якщо поява однієї з них при даному випробовуванні виключає можливість появи іншої. Події, які полягають у появі після ланцюжка *котр* букв *а* та *в*, є несумісними.

8. Дві події є *сумісними*, якщо поява однієї з них при даному випробуванні не виключає появи другої.

9. Події $A, B, C, \ldots, Z$ утворюють *повну систему подій*, якщо при виконанні випробування при комплексі умов σ хоча б одна з них має відбутись. Події, які полягають у появі після ланцюжка *котр* букв *а*, *в*. *и*, *о*, *і*, утворюють повну систему подій.

10. Дві несумісні події A та $\overline{A}$ (читається „не A"), якщо вони утворюють повну систему подій, називаються *протилежними*.



2 ЙМОВІРНІСТЬ ЕЛЕМЕНТАРНОЇ ЛІНГВІСТИЧНОЇ ПОДІЇ

Мірою можливості появи лінгвістичної події A при виконанні комплексу умов σ є ймовірність $P(A)$ цієї події. Для мовознавства важливими є три означення ймовірності: а) означення, яке ґрунтується на суб'єктивній кількісній оцінці можливості події; б) класичне означення ймовірності; в) статистичне означення ймовірності.

2.1 Суб'єктивне означення ймовірності та його використання у лінгвістиці

Якщо людина вирішує інтуїтивно оцінити ймовірність появи події C, то вона використовує сукупність знань (тезаурус) Θ відносно тих можливостей, котрі можуть сприяти або не сприяти появі події A.

Ця ймовірність може бути представлена як $P(A,\Theta)$, тобто як ймовірність події A при наявному у свідомості даної людини тезаурусі Θ. Якщо дві людини мають відносно події A однаковий тезаурус Θ, то значення ймовірностей події A для цих людей будуть однаковими. Проте, така ситуація зустрічається рідко. Частіше ймовірність однієї і тої самої події оцінюється різними людьми, виходячи з різних величин Θ,Θ'. Навіть у однієї і тої самої людини з часом величина Θ змінюється і перетворюється в Θ', отже, і його оцінки ймовірності події A у різні періоди життя є різними: $P(A,\Theta) \neq P(A,\Theta')$.

На основі використання суб'єктивних ймовірностей раніше будувалось багато мовних досліджень, а відмінності у суб'єктивних ймовірностях ставали джерелом лінгвістичних дискусій.

2.2 Класичне означення ймовірності (схема випадків) і побудова частотного словника цілісного тексту

Існують випробування, для яких ймовірності появи події можна оцінити безпосередньо з умов самого досліду. Для цього необхідно, щоб різні результати випробувань були рівноможливими.

Якщо результати випробування можна зобразити у вигляді повної системи N рівноможливих і попарно несумісних подій і якщо випадкова подія A відбувається тільки в F випадках, то *ймовірність* події A дорівнює

$$P(A) = F/N, \tag{1}$$

тобто відношенню кількості випадків, що сприяють даній події, до загальної кількості всіх випадків.

З класичного означення ймовірності випливають такі наслідки.

1. Ймовірність достовірної події дорівнює одиниці:

$$P(U)=1.$$

2. Ймовірність неможливої події дорівнює нулю:

$$P(V)=0.$$

3. Ймовірність появи випадкової події A є додатне число, яке міститься між нулем та одиницею:

$$0 \le P(A) \le 1.$$

У деяких лінгвістичних роботах, які використовують елементи теорії ймовірності, величина ймовірності виражається у процентах (0-100%).

Ґрунтуючись на класичному означенні може бути здійснена імовірнісна обробка частотних словників окремих творів або всієї творчості письменника. У цих випадках усі слововживання, які складають текст усіх творів або окремого твору, утворюють повну систему рівно можливих і попарно незалежних подій. Деяке слово (або словоформа) A, яке нас цікавить, з'являється у тексті, який досліджується, у вигляді слововживань. Звідси ймовірність того, що навмання взяте слово з нашого тексту виявиться словом (словоформою) A дорівнює $P(A)=F/N$.

2.3 Статистичне означення ймовірності. Вибірковий частотний опис тексту

Класичне означення ймовірності виявляється дуже зручним стосовно до таких дослідів, які дають скінченну кількість рівноможливих закінчень. Проте, при переході від простих прикладів до розв'язування більш складних імовірнісно-лінгвістичних задач, це означення наштовхується на непереборні труднощі.

По-перше, кількість можливих результатів практично може не бути скінченною. По-друге, стверджувати рівноможливість результатів лінгвістичного досліду буває дуже важко.

До дослідів, які не можуть бути досліджені на основі системи випадків. застосовується *статистичне означення ймовірності*.

Нехай здійснено серію з N випробувань, у кожному з яких могла з'явитись або не з'явитись подія A. Тоді *абсолютною частотою* F називається кількість появ події A, а *відносною частотою* (або просто *частотою*) $f(A)$ – відношення абсолютної частоти до загальної кількості випробувань:

$$f(A) = F/N. \qquad (2)$$

Результати окремих статистичних випробувань можуть давати помітні флуктуації. Проте, при великій кількості випробувань N статистичні флуктуації починають згладжуватись, а відносна частота f виявляє все більшу стійкість. Іншими словами, у випадкових явищах є деяка об'єктивна властивість, яка має тенденцію залишатись постійною і проявляється що ясніше при збільшенні обсягу матеріалу, що досліджується. Вказана властивість вимірюється деякою сталою величиною, яка є кількісною числовою характеристикою явища, яке вивчається. Ця стала величина і називається ймовірністю випадкової події A [будемо її, як і раніше, позначати через $P(A)$]. Експериментальними значеннями ймовірності є відносні частоти $f(A)$ досліджуваної події у певних серіях спостережень. Означена таким чином імовірність випадкової події називається *статистичною ймовірністю*.

Потрібно підкреслити, що точне числове значення статистичної ймовірності залишається, взагалі кажучи, невідомим. За числове значення ймовірності звичайно береться при великій кількості випробувань або відносна частота події A, або деяке число, близьке до неї, наприклад деяке середнє відносних частот, одержаних з декількох достатньо великих серій випробувань.

Описаний підхід має принципове значення для прикладних лінгвістичних досліджень. Не маючи, як правило, можливості дослідити всю генеральну сукупність можливих результатів, ми змушені здійснювати серію спостережень, які охоплюють деяку частинну сукупність.

3 ЙМОВІРНОСТІ СКЛАДНИХ ЛІНГВІСТИЧНИХ ПОДІЙ

3.1 Додавання ймовірностей

Мовознавця рідко цікавлять елементарні події, частіше йому доводиться мати справу зі складними лінгвістичними подіями, наприклад. із сумою елементарних подій. *Вибір правил, за допомогою яких обчислюється ймовірність складної події, визначається тим, несумісними чи сумісними є елементарні події, що утворюють складну подію.*

Ймовірність появи однієї з декількох попарно несумісних подій дорівнює сумі ймовірностей цих подій:

$$P(A_1 + A_2 + ... + A_n) = P(A_1) + P(A_2) + ... + P(A_n). \quad (3)$$

Проте, якщо дві події сумісні, то ймовірність їхньої суми обчислюється як сума ймовірностей цих подій мінус добуток ймовірностей цих подій:

$$P(A+B) = P(A) + P(B) - P(A) \cdot P(B). \quad (4)$$

Зауваження. При відшуканні ймовірності події A часто доцільно спочатку обчислити ймовірність події $\overline{A}$ (протилежної), а потім знайти шукану ймовірність за формулою

$$P(A)=1-P(\overline{A}). \tag{5}$$

Імовірність появи трьох сумісних подій обчислюється як сума ймовірностей цих подій, мінус попарні добутки імовірностей, плюс добуток імовірностей цих подій:

$$P(A+B+C)=P(A)+P(B)+P(C)-P(A)\cdot P(B)-P(B)\cdot P(C)-P(A)\cdot P(C)+ \\ +P(A)\cdot P(B)\cdot P(C). \tag{6}$$

При обчисленні ймовірності суми декількох сумісних подій використовують правило, за яким ймовірність появи хоча б однієї з декількох сумісних подій $A_1, A_2, \ldots, A_n$ дорівнює різниці між одиницею і ймовірністю одночасної появи (добутку) всіх протилежних подій. Іншими словами,

$$P(A_1 + A_2 + \ldots + A_n) = 1 - P(\overline{A}_1 \overline{A}_2 \ \ldots \ \overline{A}_n) = 1 - \prod_{i=1}^{n} (1 - P(A_i)). \qquad (7)$$

3.2 Прогнозування ймовірностей лінгвістичних подій при повторенні дослідів

Розглянуті правила широко використовуються для прогнозування подій у різного роду ймовірнісно-лінгвістичних, інженерно-лінгвістичних та інформаційних задачах. У зв'язку з цим розглянемо такий приклад.

Для побудови алгоритму роботи імовірнісного автомата, який розпізнає усну мову, доводиться обчислювати ймовірність збігу хоча б однієї із словоформ тексту, що обробляється, з відповідною лексемою, заданою у словнику автомата. Значення статистичної ймовірності появи займенника *він* дорівнює 0.0099

Припустимо, що потрібно визначити ймовірність того, що хоча б одне з двох вибраних слів тексту буде займенником *він*.

Позначимо через A першу появу займенника *він*, а через B – другу появу цього ж займенника. Події A та B сумісні, оскільки можна знайти слово *він* як у першому, так і у другому уривках. Отже, для розв'язування нашої задачі потрібно скористатись формулою (4). Враховуючи це, отримаємо

$$P(A+B)=0.0099+0.0099-0.0099\cdot 0.0099\approx 0.020\,.$$

Відзначимо очевидний факт, що значення статистичної ймовірності появи займенника *він*, обчислене за формулою (7), безумовно, співпадає з обчисленим вище (за формулою (4)):

$$P(A+B)=1-(1-0.0099)\cdot(1-0.0099)=1-(1-0.0099)^2\approx 0.020\,.$$

Тепер припустимо, що розпізнавальний автомат аналізує десять взятих навмання словоформ, і спробуємо визначити ймовірність того, що хоча б одна з цих словоформ виявиться займенником *він*. Як видно з формул (1.10) та (1.12), кількість доданків для обчислення імовірності складає $C_{10}^1 + C_{10}^2 + C_{10}^3 + ... + C_{10}^9 + C_{10}^{10} = 1023$ доданки. Очевидно, що доцільніше скористатися формулою (1.13), де A_i – подія, що полягає у появі займенника в i-й спробі. Оскільки ймовірність $P(A)$ для всіх уривків однакова, то знайдемо

$$P(A_1 + A_2 + ... + A_9 + A_{10}) = 1 - (1 - 0.0099)^{10} \approx 0.095 .$$

Як бачимо, і класичне, і суб'єктивне означення імовірності співпадають у своїх оцінках: ймовірність одержати хоча б один займенник *він* при десятикратному виборі словоформи з тексту є помітно вищою, ніж ймовірність одержати його при однократному або двократному виборі.

3.3 Залежні лінгвістичні події та умовні ймовірності

Досі ми мали справу з *незалежними* подіями, тобто такими подіями, ймовірність появи яких не залежала від ймовірності появи іншої лінгвістичної події – такі ймовірності називаються *безумовними*. Проте, мовознавство порівняно рідко має справу з незалежними подіями. Звичайно мова йде про залежні події та умовні ймовірності: навіть ймовірності появи букв, фонем, складів, морфем тощо є умовними, оскільки залежать від позиції цих лінгвістичних об'єктів у слові, словосполученні і реченні.

Розглянемо співвідношення залежних і незалежних лінгвістичних подій, а також безумовних і умовних імовірностей на прикладі штучного лінгвістичного досліду.

Словоформа *мамам* (давальний відмінок множини від *мама*) складена з букв розрізної абетки. Картки з буквами цієї словоформи покладені в урну. Здійснюється випробування, яке полягає у витяганні картки з буквою і поверненні її в урну. Подією B вважається витягання букви *м* у першому випробуванні (тоді $\overline{B}$ буде витягання з урни не *м*, тобто у цьому прикладі це означає витягання букви *а*), подією A – витягання букви *а* у другому випробуванні (тоді $\overline{A}$ буде витягання з урни не *а*, тобто букви *м*). Оскільки витягнена в перший раз буква повертається в урну, то перед другим дослідом кількість букв в урні не зміниться. Тому ймовірність події A є безумовною, оскільки вона не залежить від того, чи була витягнена до цього з урни буква *м* (подія B) чи буква *а* (подія $\overline{B}$), і залишається рівною $2/5$. Безумовною є і ймовірність події B.

Якщо змінити умови досліду і не повертати витягнену букву назад до урни, то ймовірності одержати у другому, третьому і наступних випробуваннях букву *а* або *м* будуть істотно залежати від того, які букви були витягнені перед цим з урни.

Нехай результатом першого випробування була буква *м*; тоді ймовірність витягнути у другому випробуванні букву *а* складе $2/4 = 1/2$. У тому ж випадку, коли в результаті першого досліду одержана була буква *а* (подія $\overline{B}$), ймовірність витягнути другий раз букву *а* дорівнює $1/4$. Аналогічна ситуація виникає при визначенні ймовірності появи букви *м* (подія $\overline{A}$) у другому витягуванні за умови, що у перший раз була витягнена буква *м* (B) або *а* (подія $\overline{B}$). Іншими словами, події A та B є залежними, а їхні ймовірності – умовними.

Умовна ймовірність події A за умови, що відбулась подія B, позначається через $P(A/B)$. Так, у розглянутому прикладі

$$P(A/B) = 1/2,\ P(\overline{A}/B) = 1/2,\ P(A/\overline{B}) = 1/4,\ P(\overline{A}/\overline{B}) = 3/4.$$

Умовна ймовірність події A, обчислена за умов, що відбулось декілька подій $B_1, B_2, ..., B_k$ позначається через $P(A/B_1B_2\ \ldots\ B_k)$.

Величина умовної ймовірності завжди міститься в тому ж проміжку, що і величина абсолютної ймовірності, тобто

$$0 \le P\ (A/B_1B_2\ \ldots\ B_k) \le 1.$$

3.4 Правило множення ймовірностей і обчислення ймовірностей мовних елементів

Кожний текст або його частину можна розглядати як сумісну появу деякої лінійної послідовності лінгвістичних подій – сумісну появу ланцюжка словоформ, послідовності складів, ланцюжків фонем або букв. Визначення ймовірностей появи цих ланцюжків ґрунтується на *теоремі множення ймовірностей*.

Ймовірність сумісної появи двох подій дорівнює добутку ймовірності першої події на умовну ймовірність другої, обчислену за умови, що перша подія відбулась:

$$P(AB) = P(A)P(B/A). \qquad (8)$$

Наслідки.

1. Застосуємо формулу (8) до події BA:

$$P(BA)=P(B)P(A/B),$$

і, оскільки події AB та BA не відрізняються, то

$$P(AB)=P(B)P(A/B). \tag{9}$$

Порівнюючи формули (8) та (9) одержуємо, що

$$P(A)P(B/A)=P(B)P(A/B). \tag{10}$$

2. Якщо подія A не залежить від B, то і подія B не залежить від A.

Для незалежних подій теорема множення ймовірностей спрощується: ймовірність добутку двох незалежних випадкових подій дорівнює добутку їх безумовних ймовірностей:

$$P(AB)=P(A)P(B). \tag{11}$$

3. Якщо події A та B незалежні, то незалежні також і пари подій $\{\overline{A}, B\}$, $\{A, \overline{B}\}$, $\{\overline{A}, \overline{B}\}$.

4. Ймовірність добутку залежних подій A, B, C дорівнює добутку ймовірності однієї з них на умовну ймовірність другої, обчислену за умови, що перша подія відбулась, і на умовну ймовірність третьої, обчислену за умови, що дві попередні події відбулись:

$$P(ABC) = P(A)P(B/A)P(C/AB). \tag{12}$$

Узагальнюючи цей наслідок на n залежних подій A_1, A_2, ... A_n, одержуємо

$$P\left(\prod_{i=1}^{n} A_i\right) = P(A_1) \cdot P(A_2/A_1) \cdot P(A_3/A_1A_2) \cdot P\left(A_n/\prod_{i=1}^{n} A_i\right). \tag{13}$$

Використовуючи описану у цьому пункті теорію, можна обчислювати ймовірності появи у письмових текстах різних ланцюжків букв.

Розглянемо такий приклад. Нехай відносна частота букви *я* на початку слова 0.035, а відносна частота пробілу Δ у тексті 0.174. Тоді ймовірність появи ланцюжка $\Delta я$ дорівнює

$$P(\Delta я) = P(\Delta) \cdot P(я / \Delta) = 0.174 \cdot 0.035 = 0.006 = 0.6\%.$$

Нехай імовірність появи пробілу Δ та букви *и* після ланцюжка $\Delta я$ складає, відповідно, 0.701 та 0.001. Щоб визначити ймовірність появи слова *я*, утворимо ланцюжок $\Delta я \Delta$, для якого

$$P(\Delta я \Delta) = P(\Delta) \cdot P(я / \Delta) \cdot P(\Delta / \Delta я) = 0.174 \cdot 0.035 \cdot 0.701 \approx 0.00427 \approx 0.4\%.$$

Тепер розрахуємо ймовірність появи морфеми *япон*. Для цього формуємо ланцюжок $\Delta япон$, тоді

$$P(\Delta япон) = P(\Delta) \cdot P(я / \Delta) \cdot P(п / \Delta я) \cdot P(о / \Delta яп) \cdot P(н / \Delta япо).$$

Із аналізу словників можна зробити висновок, що після ланцюжка *Δяп* єдино можливою буде діграма *он*. Звідси випливає, що появи тут букв *о* та *н* є достовірними подіями, умовна ймовірність яких дорівнює одиниці. Таким чином,

$$P(\Delta япон) = 0.174 \cdot 0.035 \cdot 0.001 \cdot 1 \cdot 1 \approx 0.00006 = 0.006\%.$$

3.5 Визначення загальної ймовірності лінгвістичної події за формулою повної ймовірності

Якщо лінгвістична подія A може відбутись разом з однією і тільки однією з n несумісних подій H_1, H_2, ..., H_n, які утворюють повну групу подій, то для визначення ймовірності події A використовується формула *повної ймовірності*:

$$P(A)=\sum_{i=1}^{n} P(H_i)\cdot P(A/H_i). \qquad (14)$$

Несумісні події $H_1, H_2, ..., H_n$ називаються *гіпотезами*. Таким чином, ймовірність події A дорівнює сумі добутків ймовірності кожної гіпотези на ймовірність події при здійсненні цієї гіпотези.

Формула повної імовірності використовується для обчислення загальної ймовірності лінгвістичної події за умови, що відомі її ймовірності у вузькотематичних вибірках.

Нехай, наприклад, є англійський науково-технічний текст загальною довжиною в 400 тис. слововживань (близько тисячі стандартних сторінок). За тематикою цей текст розпадається на такі чотири вибірки різної довжини:

1) радіоелектроніка – 200 тис. слововживань (прибл. 500 с.),

2) автомобілебудування – 100 тис. слововживань (прибл. 250 с.),

3) корабельні механізми – 50 тис. слововживань (прибл. 125 с.),

4) будівельні матеріали – 50 тис. слововживань (прибл. 125 с.).

Словоформа *are* – множина дійсного часу дієслова *to be* (*бути*) – вжита у 1-й вибірці 1610, у 2-й – 1273, у 3-й – 469 і у 4-й – 346 разів. Аналогічно словоформа *machine* (*машина, механізм*) зустрілась у 1-й вибірці 98, у 2-й – 57, у 3-й – 9 і у 4-й – 19 разів.

Для простоти сприйняття організуємо умову задачі у табличному вигляді:

№	Тематика вибірки	Об'єм вибірки	*to be*	*machine*
1	Радіоелектроніка	200000	1610	98
2	Автомобілебудування	100000	1273	57
3	Корабельні механізми	50000	469	9
4	Будівельні матеріали	50000	346	19

Потрібно визначити ймовірність того, що взяте навмання з нашого тексту слово буде: а) словоформою *are*; б) словоформою *machine*.

Для цього вважатимемо появу словоформи *are* подією A, а появу *machine* – подією B. Розглянемо також такі чотири гіпотези: H_1 – приналежність словоформи до текстів з радіоелектроніки, H_2 – до текстів з автомобілебудування, H_3 – до текстів з корабельних механізмів, H_4 – до текстів з будівельних матеріалів.

Вважаючи частки вказаних текстів у загальній вибірці ймовірностями наших гіпотез, обчислюємо:

$$P(H_1) = 200000/400000 = 0.5;\ P(H_2) = 100000/400000 = 0.25;$$

$$P(H_3) = 50000/400000 = 0.125.$$

Умовні ймовірності події A (поява дієслова *are*) за цих гіпотез відповідно дорівнюють:

$$P(A/H_1) = 1610/200000 = 0.008;\ P(A/H_2) = 1273/100000 = 0.012;$$

$$P(A/H_3) = 469/50000 = 0.009;\ P(A/H_4) = 346/50000 = 0.007.$$

Застосовуючи формулу повної ймовірності, визначаємо, що ймовірність вибрати навмання з даного тексту словоформу *are* дорівнює

$$P(A)=P(H_1)\cdot P(A/H_1)+P(H_2)\cdot P(A/H_2)+P(H_3)\cdot P(A/H_3)+P(H_4)\cdot P(A/H_4)=$$
$$=0.5\cdot 0.008+0.25\cdot 0.012+0.125\cdot 0.009+0.125\cdot 0.007\approx 0.009=0.9\%.$$

Аналогічно обчислюємо умовні ймовірності події B (поява *machine*):

$$P(B/H_1)=98/200000=0.0005;\ P(B/H_2)=57/100000=0.0006;$$
$$P(B/H_3)=9/50000=0.0002;\ P(B/H_4)=19/50000=0.0004.$$

За формулою повної ймовірності одержуємо, що ймовірність дістати з даного тексту словоформу *machine* складає

$$P(B)=0.5\cdot 0.0005+0.25\cdot 0.0006+0.125\cdot 0.0002+0.125\cdot 0.0004=0.000475\approx 0.048\%$$

3.6 Апріорні та апостеріорні ймовірності. Вимірювання ймовірностей лінгвістичних гіпотез

Досі ми мали справу з так званими *апріорними* ймовірностями лінгвістичних подій. Ці апріорні ймовірності встановлювались інтуїтивно-емпірично або теоретично до здійснення досліду, виходячи з наших знань про умови σ цього досліду. Наші відомості про умови досліду звичайно неповні, тому апріорні ймовірності є ймовірностями деяких лінгвістичних гіпотез $H_1, H_2, ..., H_n$ про результат експерименту.

Результат експерименту, як правило, змушує здійснити переоцінку наших гіпотез і надати їм нові – *апостеріорні* ймовірності. Визначення апостеріорних імовірностей здійснюється так.

Нехай апріорні ймовірності гіпотез до досліду відповідно дорівнюють $P(H_1)$, $P(H_2)$,..., $P(H_n)$, а в результаті досліду з'явилась подія A. Необхідно визначити, як потрібно змінити ймовірності наших лінгвістичних гіпотез у зв'язку з появою події A.

За теоремою множення ймовірностей для залежних подій, ймовірність сумісної появи події A і гіпотези H_1 складає

$$P(AH_i) = P(A) \cdot P(H_i / A) = P(H_i) \cdot P(A / H_i). \quad (15)$$

Звідси випливає, що

$$P(H_i / A) = \frac{P(H_i) \cdot P(A / H_i)}{P(A)}. \quad (16)$$

Підставимо для $P(A)$ його вираз з формули повної ймовірності (14) і одержимо

$$P(H_i / A) = \frac{P(H_i) \cdot P(A / H_i)}{\sum_{j=1}^{n} P(H_j) \cdot P(A / H_j)}. \quad (17)$$

Вираз (17) називається *формулою Байєса*, або *формулою ймовірностей гіпотез*.

Щоб показати, як за допомогою формули Байєса вимірюються ймовірності лінгвістичних гіпотез, розглянемо знову виявлення в англійському науково-технічному тексті словоформ *are* та *machine* (див. п. 3.5).

Припустимо, що перша взята навмання з англійського науково-технічного тексту словоформа виявилась дієсловом *are* (подія A). Необхідно знайти ймовірність того, що ця словоформа взята: а) із тексту з радіоелектроніки (H_1); б) із тексту з автомобілебудування (H_2); в) із тексту з корабельних механізмів (H_3); г) із тексту з будівельних матеріалів (H_4).

Імовірності того, що взята словоформа належить до тої чи іншої тематичної вибірки, є апостеріорними ймовірностями гіпотез, – точніше, умовними ймовірностями цих гіпотез за умови, що відбулась подія A. Використовуючи формулу (15), одержимо

$$P(H_1/A)=\frac{P(H_1)\cdot P(A/H_1)}{P(H_1)\cdot P(A/H_1)+P(H_2)\cdot P(A/H_2)+P(H_3)\cdot P(A/H_3)+P(H_4)\cdot P(A/H_4)}=$$

$$=\frac{0.5\cdot 0.008}{0.5\cdot 0.008+0.25\cdot 0.012+0.125\cdot 0.009+0.125\cdot 0.007}\approx 0.444.$$

Аналогічно,

$$P(H_2/A)=0.333,\ P(H_3/A)=0.128,\ P(H_4/A)=0.095.$$

Використовуючи наведені вище дані, визначимо апостеріорні ймовірності гіпотез H_1, H_2, H_3, H_4 за умови, що з тексту двічі брались дві словоформи, причому обидва рази цими словоформами виявилось дієслово *are*. Експеримент будувався таким чином, що обидві словоформи могли бути взяті тільки із однієї тематичної вибірки.

ЗАВДАННЯ

ДЛЯ ЛАБОРАТОРНОЇ РОБОТИ №2

1. Яка імовірність того, що слово, що починається з тих же трьох букв, що і прізвище студента, четвертою буквою матиме букву "а" (розглядати лише слова, допущені нормами української мови).

2. Визначити ймовірність того, що хоча б одне з трьох вибраних слів тексту буде займенником *він*. Значення статистичної ймовірності появи займенника *він* дорівнює 0.0099. Використати формули (6), (7) та порівняти результати.
3. Обчислити імовірність появи морфем, що починаються з пробілу та таких трьох букв, що і прізвище студента. Відносну частоту появи кожної букви знайти з допомогою словника. Імовірність появи пробілу в тексті 0.174.

4. Визначити частоти букв в українських літературних текстах (на матеріалах довільного поетичного уривку довжиною не менше 30 слів).
5. Визначити частоти перших букв в українських літературних текстах (на матеріалах довільного поетичного уривку довжиною не менше 100 слів).

6. Нехай є український науково-технічний текст загальною довжиною 300 тис слововживань (750 стандартних сторінок). За тематикою цей текст розподілений у трьох вибірках різної довжини:

1) інформатика 200 тис. слововживань (500 с.);

2) медицина 80 тис. слововживань (200 с.);

3) логістика 20 тис. слововживань (50 с.).

Словоформа *комп'ютер* використана у першій вибірці 450, у другій – 8, у третій – 10 разів. Аналогічно, словоформа *аналіз* зустрілась у першій вибірці 5, у другій – 40, у третій – 8 разів.

a. Обчислити ймовірність того, що навмання взяте із нашого науково-технічного тексту слововживання буде: а) словоформою *комп'ютер*; б) словоформою *аналіз*.

b. В умовах задачі 6 обчислити ймовірності того, що навмання взята словоформа *комп'ютер* належить до тої чи іншої тематичної вибірки.

c. В умовах задачі 6 обчислити ймовірності того, що навмання взята словоформа аналіз належить до тої чи іншої тематичної вибірки.