МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
«ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»

С. В. Петрасова, Н. Ф. Хайрова

СУЧАСНІ ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ В ЛІНГВІСТИЦІ

Навчальний посібник
з курсу «Актуальні проблеми прикладної та теоретичної лінгвістики»
для студентів спеціальності «Прикладна та комп'ютерна лінгвістика»

Рекомендовано Вченою радою НТУ «ХПІ»

Харків
2020

УДК 004.912(075.8)
П30

Рецензенти:

С. Г. Удовенко, д-р. техн. наук, професор, Харківський національний економічний університет імені Семена Кузнеця;

І. В. Шостак, д-р. техн. наук, професор, Національний аерокосмічний університет ім. М. Є. Жуковського «Харківський авіаційний інститут»

Рекомендовано Вченою радою НТУ «ХПІ» як навчальний посібник для студентів спеціальності «Прикладна та комп'ютерна лінгвістика», протокол № 1 від 31.01.2020 р.

Петрасова С. В.

П30 Сучасні інформаційні технології в лінгвістиці : навч. посіб. / С. В. Петрасова, Н. Ф. Хайрова. Харків : ФОП Панов А.М., 2020. 124 с.

ISBN 978-617-7859-31-3

У навчальному посібнику наведено матеріал, що охоплює широке коло питань, пов'язаних з використанням сучасних інформаційних технологій при вирішенні завдань прикладної та теоретичної лінгвістики. Викладення теоретичних положень супроводжується прикладами застосування інформаційних технологій, що істотно сприяє розумінню вирішення професійних лінгвістичних задач.

Призначено для студентів спеціальності «Прикладна та комп'ютерна лінгвістика».

Іл. 17. Табл. 5. Бібліогр. 37 назв.

УДК 004.912(075.8)

ISBN 978-617-7859-31-3

ВСТУП

Курс *«Актуальні проблеми прикладної та теоретичної лінгвістики»* орієнтовано на формування навичок та знань в області прикладної та теоретичної лінгвістики, ознайомлення студентів-лінгвістів з сучасними задачами та напрямками досліджень прикладної лінгвістики, міждисциплінарними зв'язками лінгвістики з іншими науками.

Навчальний посібник акцентує увагу на актуальних аспектах прикладної лінгвістики та її зв'язку з сучасними інформаційними технологіями.

Структура навчального видання складається з 6 розділів, кожен з яких включає теоретичну частину, творчі завдання та списки бібліографічних джерел. *Теоретична частина* містить визначення ключових понять, аналіз методів та технологій, що використовуються при вирішенні актуальних проблем з представлених напрямків лінгвістики. *Творчі завдання* передбачають самостійну роботу студентів над проблемами, окресленими в розділах. Вирішення завдань сприятиме формуванню вмінь та навичок експериментальної перевірки теоретичних положень. До кожного розділу додається *список літератури*, який містить праці відомих мовознавців. Список джерел сприятиме розширенню сфер наукового пошуку.

У першому розділі систематизовано та проаналізовано головні напрямки лінгвістики, подано ключові завдання прикладної лінгвістики з огляду на сучасні інформаційні технології. Міститься аналіз методологічних засад лінгвістичних досліджень, які детально описано в наступних розділах.

У другому розділі підлягають аналізу ключові поняття структурної та математичної лінгвістики, особливу увагу зосереджено на проблемі моделювання природної мови.

Третій розділ присвячено актуальним проблемам комп'ютерної лінгвістики. Викладено методи та технології, що використовуються для вирішення цих проблем.

Четвертий розділ розкриває прикладні аспекти корпусної лінгвістики, містить аналіз сучасних корпусних технологій.

У п'ятому розділі подано аналіз комп'ютерних технологій в лексикографії, розглянуто класифікацію словників та їхні структурні елементи.

Шостий розділ присвячено міждисциплінарним дослідженням у лінгвістиці. У межах розділу здійснено аналіз ключових проблем кожного напрямку, розглядаються прикладні методи та отримані результати досліджень.

Наприкінці навчального посібника наведено додатки. *Додаток А* містить теми рефератів, які відповідають програмі навчальної дисципліни. *Додаток Б* надає опис структури та правила оформлення реферату. У *Додатку В* наведено зразок оформлення титульного аркуша реферату.

Виклад теоретичних питань і положень у посібнику є логічно структурованим та широко ілюструється прикладами застосування сучасних інформаційних технологій. Навчальний посібник повністю відповідає вимогам освітньо-професійної програми підготовки магістрів прикладної лінгвістики.

РОЗДІЛ 1
РОЛЬ ПРИКЛАДНОЇ ТА ТЕОРЕТИЧНОЇ ЛІНГВІСТИКИ В ІНФОРМАЦІЙНОМУ СУСПІЛЬСТВІ

1.1. Структура лінгвістики як науки.
1.2. Сучасні напрямки розвитку прикладної лінгвістики.
1.3. Лінгвістичні методи досліджень.

1.1. Структура лінгвістики як науки

Лінгвістика – наука про природну мову і всі мови світу як індивідуальні її представники [1, с. 7].

Макролінгвістика включає два взаємопов'язані розділи: теоретичну лінгвістику та прикладну лінгвістику. У теоретичній лінгвістиці сформувалися чотири основні розділи: внутрішня лінгвістика, зовнішня лінгвістика, інтерлінгвістика і «проміжна» лінгвістика (рис. 1.1).

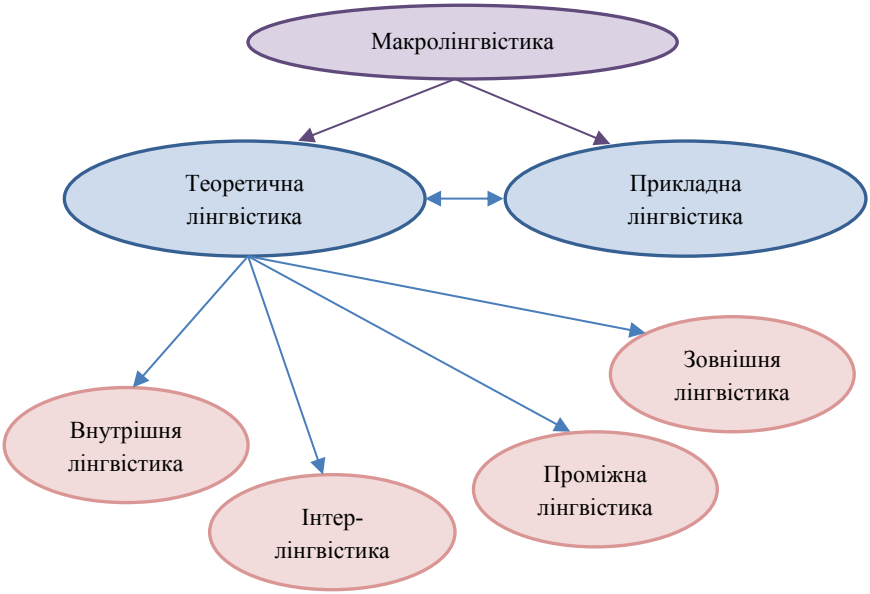

Рисунок 1.1 – Структура лінгвістики

Внутрішня лінгвістика вивчає функціонально-прагматичні властивості мовних одиниць. Розподіл внутрішньої лінгвістики проводиться залежно від рівня мовних знаків, які виступають об'єктом дослідження:

1. *Фонетика* – розділ мовознавства, що вивчає звукову сторону мови, включає:

1) акустичну;

2) артикуляційну;

3) експериментальну;

4) функціональну фонетику.

2. *Фонологія* вивчає фонологічну систему мови і включає:

1) сегментну;

2) суперсегментну фонологію.

3. *Лексикологія* вивчає словниковий склад, лексику мови в її сучасному та історичному розвитку, а також включає:

1) семасіологію – розділ мовознавства, що займається лексичною семантикою, тобто значеннями слів і словосполучень, які використовуються для номінації окремих предметів і явищ;

2) ономасіологію – розділ мовознавства, який займається вивченням принципів називання (номінації).

4. *Фразеологія* вивчає фразеологічний склад мови в її сучасному стані та історичному розвитку.

5. *Граматика* вивчає граматичну будову мови в її граматичних категоріях, граматичних одиницях і граматичних формах та включає:

1) морфологію – розділ граматики, що вивчає закономірності функціонування і розвитку морфологічної системи мови, в яку входять:

– формоутворення (закономірності творення форм слів);

– теорія частин мови (класів слів, що виділяються на підставі спільності їхніх морфологічних, синтаксичних і семантичних властивостей);

2) морфеміку – розділ граматики, що вивчає морфемний склад мови: типи і структуру морфем, їхні відношення один до одного і до слова в цілому;

3) словотворення – розділ граматики, що вивчає способи творення, функціонування, будови та класифікації похідних і складних слів;

4) синтаксис – розділ граматики, що вивчає процеси породження мовлення: сполучуваність і порядок розташування слів всередині речення, а також загальні властивості речення як самостійної одиниці мови (та мовлення).

6. *Семантика* – розділ мовознавства, що вивчає зміст, інформацію, передану мовою або будь-якою її одиницею (словом, граматичною формою слова, словосполученням, реченням).

7. *Лінгвістика тексту* – розділ мовознавства, в якому вивчаються правила побудови зв'язного тексту і його смислові (текстові) категорії, що виражаються цими правилами.

8. *Лінгвосеміотика* – галузь мовознавства, в якій вивчається специфіка мовних знаків і специфіка мови як знакової системи, і формуються знакові теорії мови.

9. *Менталінгвістика* – галузь мовознавства, в якій вивчається взаємовідношення мови і мислення, мови і свідомості, а також концепції вербалізації мовних процесів. Мова розглядається як найважливіший засіб спілкування людей і засіб вираження їхніх думок і почуттів. Менталінгвістика пов'язана з психолінгвістикою, нейролінгвістикою і когнітивною лінгвістикою.

10. *Комунікативна лінгвістика* – галузь мовознавства, в якій вивчається взаємодія субстанціональних і функціональних властивостей у ситуації реального спілкування в усній і письмовій формах.

11. *Лінгвістична типологія* – лінгвістична галузь, в якій здійснюється порівняльне вивчення структурних і функціональних властивостей мов незалежно від характеру генетичних відношень між ними та ареальних об'єднань, а на основі мовного типу.

12. *Лінгвістика універсалій* – розділ мовознавства, в якому акцент робиться на властивостях, притаманних усім мовам або більшості з них, тобто верифікації та інтерпретації лінгвістичних універсалій.

Зовнішня лінгвістика – внутрішня лінгвістика, яка використовує для вивчення мови методи суміжних наук, як правило, спільно з лінгвістичними методами.

Диференціація зовнішньої лінгвістики:

1. *Психолінгвістика* – лінгвістична наука, що склалася на стику психології і лінгвістики, що вивчає процеси творення мовлення, а також сприйняття та формування мови в їх співвіднесеності з системою мови. Як гібридна наука психолінгвістика за предметом дослідження близька до лінгвістики, а за методами – до психології.

2. *Нейролінгвістика* – лінгвістична дисципліна, що виникла на стику неврології (нейрофізіології) і лінгвістики та вивчає психо-фізіологічний механізм мовного відображення дійсності. В останні десятиліття актуальними стали дослідження зворотного зв'язку – вплив мови на протікання фізіологічних процесів.

3. *Ареальна лінгвістика* – лінгвістична наука, що досліджує за допомогою методів лінгвістичної географії поширення мовних явищ (мов і діалектів) у просторовій протяжності і міжмовній (междіалектній) взаємодії.

4. *Паралінгвістика* – мовознавча дисципліна, що вивчає невербальні, немовні засоби (кінетичні, графічні), що включені в мовне спілкування і передають разом з вербальними засобами смислову інформацію.

5. *Математична лінгвістика* – суміжна дисципліна, яка виникла на стику математики і лінгвістики, предметом якої є розробка формального апарату для опису будови природних і деяких штучних мов; теорія способів опису синтаксичної структури мов.

6. *Лінгвостатистика* – лінгвістична дисципліна, що розробляє способи застосування статистичних методів, які базуються на статистико-імовірнісних закономірностях, у мовознавстві насамперед для вивчення процесів мови.

7. *Палеографія* – історико-філологічна дисципліна, що вивчає створення знаків писемності та їхній розвиток, а також закономірності розвитку писемності.

8. *Лінгвістична палеонтологія* – галузь лінгвістики, в якій за допомогою відповідних прийомів виявляються найбільш архаїчні елементи мови, що зберігаються в історії.

9. *Когнітивна лінгвістика* – одна з міждисциплінарних наук, у центрі уваги якої знаходиться мова як загальний когнітивний (пізнавальний) механізм, як когнітивний інструмент – система знаків, що відіграють роль у репрезентації (кодуванні) та у трансформуванні інформації.

10. *Політична лінгвістика* – наука, що виникла на перетині лінгвістики з політологією та враховує також досягнення етнології, соціальної психології, соціології та інших гуманітарних наук, активно вбирає в себе евристики дискурс-аналізу та когнітивної методології, предметом дослідження якої є політична комунікація.

11. *Лінгвокультурологія* – дисципліна, що виникла на стику мовознавства (психолінгвістики, соціолінгвістики, етнолінгвістики, лінгвокраїнознавства, прагмалінгвістики та когнітивної лінгвістики) і культурології, яка зосередила свою увагу на взаємодії мови, як транслятора культурної інформації, культури з її установками і уподобаннями, та людини, яка створює цю культуру, користуючись мовою.

Інтерлінгвістика – галузь мовознавства, що вивчає міжнародні мови як засіб міжмовного спілкування людей різних держав, в якій основна увага звертається на:

– процеси виникнення *природних міжнародних мов* (наприклад, лінгва франка, койне, піджини, для яких функція міжмовного спілкування є реальною, але вторинною за відношенням до ролі етнічної мови);

– процеси створення і функціонування *міжнародних штучних мов* (наприклад, есперанто, інтерлінгва):

1) в опозиції до природних мов знаходяться так звані планові мови – *неспеціалізовані штучні міжнародні мови*, які розшифровуються як «міжнародна + штучна + комунікативно реалізована»;

2) ще одна група міжнародних мов – *спеціалізовані штучні мови*, створювані для точної передачі різноманітної спеціальної (професійної) інформації, наприклад, так звана «біологічна латина».

Четвертий розділ у теоретичній лінгвістиці, визначений як ***проміжна лінгвістика***, вивчає соціальні та просторово-часові умови функціонування мови:

1. *Діалектологія* – традиційний розділ мовознавства, що вивчає місцеві, територіальні різновиди мови – діалекти. Діалектологія підрозділяється на:

1) описову діалектологію, яка вивчає особливості територіальних діалектів, що встановлює характер їхніх фонетичних, граматичних та лексичних рис і певних немовних (соціально-історичних) чинників, які відрізняють одні діалекти від інших;

2) історичну діалектологію, яка здійснює реконструкцію історії різних діалектів, історію виникнення, розвитку або втрати діалектних особливостей і встановлення діалектних різновидів будь-якої мови (мов) протягом всього її існування і розвитку.

2. *Лінгвістична географія* – розділ мовознавства, що виділився з діалектології і вивчає територіальне поширення мовних явищ у взаємодії з ареальною лінгвістикою, що дає можливість на основі порівняльного вивчення ізоглос отримати важливі відомості для вивчення історії мов і діалектів, встановити їхні історичні зв'язки.

3. *Соціолінгвістика* – міждисциплінарна наука, що виникла на стику мовознавства, соціології, соціальної психології та етнографії і представляє собою реалізацію на сучасному етапі лінгвістичних і соціологічних досліджень принципів і процедур соціологічного напрямку в мовознавстві. Соціолінгвістика вивчає широкий комплекс проблем, пов'язаних з соціальною природою мови, її суспільними функціями, механізмом впливу соціальних чинників на мову і тією роллю, яку відіграє мова в житті суспільства.

4. *Компаративістика* – галузь мовознавства, об'єктом якої є споріднені мови, тобто генетично пов'язані мови, які порівнюються з метою встановлення співвідношення між ними і опису їхньої еволюції в часі і просторі на базі порівняльно-історичного методу, що базується на принципах порівняння та історизму.

5. *Контрастивна лінгвістика* – галузь мовознавства, яка займається порівняльним вивченням двох, рідше декількох мов для виявлення їхніх подібностей і відмінностей (або тільки відмінностей) на всіх рівнях мовної структури, як правило, на синхронному зрізі мови з використанням порівняльного, контрастивного методу.

6. *Контактна лінгвістика* – міждисциплінарна мовознавча наука, що оформляється на стику етнолінгвістики, соціолінгвістики, психолінгвістики, компаративістики, контрастивної лінгвістики та інтерлінгвістики, яка вивчає різноманітні механізми мовних контактів (в структурному, соціально-культурному, етноісторичному, етнокультурному і антропологічному аспектах) двох або декількох мов при певних соціально-історичних умовах, а також вивчає явища, що виникли в результаті цих контактів, зокрема – контактні елементи.

7. *Етимологія* – розділ мовознавства, пов'язаний з діалектологією і історичною лексикологією, в якому вивчається походження слів на основі реконструкцій первинних форм і значень за допомогою етимологічного аналізу на основі порівняльно-історичного методу.

8. *Ономастика* – розділ мовознавства, пов'язаний з історією, етнографією, текстологією і вивчає історію виникнення назв і мотиви їхньої номінації.

Ономастика включає:

1) антропоніміку – розділ ономастики, що вивчає антропоніми – власні назви людей;

2) топоніміку – розділ ономастики, що вивчає топоніми – географічні назви;

3) етноніміку – розділ ономастики, що вивчає походження, поширення і функціонування етнонімів – слів, які називають різні види етнічних спільнот (націй, народів, народностей, племен, родів і т.п.) і лінгвонімів (назв мов і діалектів).

9. *Лінгвістична стилістика* – комплексна наукова мовознавча дисципліна, яка вивчає виражальні засоби і можливості мови та закономірності функціонування мови в різних сферах суспільної діяльності і ситуаціях

спілкування. Предметом стилістики є стиль у всіх його значеннях, проте завдання сучасної стилістики ширше, оскільки вона досліджує і питання формування стилів у зв'язку з історією літературної мови, і мову художньої літератури, і проблеми організації тексту, і різні жанри спілкування, виходячи з урахуванням структури мовного акту і його успішності, ефективності.

Такі основні розділи теоретичної лінгвістики, з якою у двосторонніх зв'язках знаходиться прикладна лінгвістика: з одного боку, спільні методи вирішення завдань у прикладній лінгвістиці розробляються теоретичним мовознавством, а з іншого, – огляд проблематики будь-якого прикладного напрямку виявляє в ньому значну теоретичну складову [2, с. 351–361].

1.2. Сучасні напрямки розвитку прикладної лінгвістики

Прикладна лінгвістика – напрям у мовознавстві, що займається розробкою методів вирішення практичних завдань, пов'язаних з використанням мови [3, с. 6].

Прикладна лінгвістика є комплексною науковою дисципліною, яка застосовує лінгвістичні знання в різних ситуаціях для вирішення різного роду практичних завдань (таких, наприклад, як машинний переклад, технічна комунікація, розпізнавання і синтез мови, інформаційний пошук та ін.).

Американське лінгвістичне суспільство визначає прикладну лінгвістику як: «Applied linguistics refers to a broad range of activities which involve solving some language-related problem or addressing some language-related concern» [1, с. 12].

Основні напрямки прикладного мовознавства:

- лінгвістичне дешифрування;
- лінгвістична (стилістична) діагностика;
- лінгвістична криміналістика;
- нормування мови та мовна політика;
- комп'ютерна лінгвістика;

– автоматизовані навчальні системи;
– лексикографія;
– автоматичне розпізнавання мови;
– представлення знань;
– машинний переклад;
– лінгвістичне забезпечення інформаційно-пошукових систем;
– автоматичне реферування і анотування;
– корпусна лінгвістика;
– проектування баз даних і баз знань;
– прикладні аспекти квантитативної лінгвістики;
– нейролінгвістичне моделювання;
– психолінгвістичне моделювання;
– прикладні аспекти політичної лінгвістики;
– лінгвістичні аспекти штучного інтелекту та ін. [2, с. 362–363].

Сучасна прикладна лінгвістика активно розвивається і вирішує багато проблем обробки усного та писемного мовлення, витягу та використання інформації з текстів, підвищує ефективність комунікації, у тому числі у сфері інформаційних і комп'ютерних технологій.

Прикладна лінгвістика являє собою ту сферу, де реально проводяться лінгвістичні експерименти, що мають на меті верифікацію положень теорії мовознавства і перевірку ефективності лінгвістичних продуктів, що створюються розробниками [1, с. 13].

1.3. Лінгвістичні методи досліджень

Метод – у широкому сенсі, спосіб організації пізнавальної і дослідницької діяльності вченого з метою вивчення явищ і закономірностей певного об'єкта науки; у вузькому сенсі, система процедур вивчення об'єкта дослідження і / або перевірки отриманих результатів.

За диференціацією методів як способів дослідницької діяльності їх можна розділити на загальні, які є дослідницькими операціями з явищами

об’єктів будь-якої науки, і лінгвістичні, що застосовуються саме в лінгвістиці з метою вивчення її об’єкта та предмета.

До ***загальних методів*** належать спостереження, індукція, дедукція, гіпотеза, аналіз, синтез, таксономія, порівняння, формалізація, ідеалізація, експеримент, фальсифікація, моделювання і т.п.

Спостереження є цілеспрямованим сприйняттям, обумовленим завданням пізнавальної діяльності, головною умовою якого є максимальна об’єктивність, контрольована шляхом повторного спостереження або застосуванням інших методів. Одним з різновидів спостереження є інтроспекція – самоспостереження. У лінгвістиці самоспостереження на підставі інтуїції дослідника досить поширений і найпростіший шлях перевірки гіпотез, однак головним недоліком інтроспекції є її суб’єктивність, що може призвести до неправильної оцінки досліджуваних фактів.

Індукція розглядається як метод узагальнення результатів окремих конкретних спостережень і експериментів (шлях від конкретного, одиничного до загального). У мовознавчих дослідженнях індуктивні методи застосовуються при створенні різних класифікацій мовних явищ, вивченні універсалій, типології мов.

Дедукція є методом переходу від загальних тверджень до конкретного висновку, знання про одиничне, які знаходяться в логічних відношеннях прямування. У лінгвістиці дедуктивні методи використовуються для встановлення статусу різних мовних одиниць, належності їх до певної частини мови, категорії за умови визначення загальної кваліфікації ознак статусу, частини мови, розряду, категорії і т.д.

Гіпотеза є побудовою припущення щодо ознак досліджуваного об’єкта, його закономірностей, способів існування і т.д. Вона є варіантом можливого вирішення проблеми, що висувається апріорно і вимагає перевірки і підтвердження. Подальший аналіз проблеми підтверджує гіпотезу, перетворюючи її в наукову теорію, або спростовує її, хоча існують гіпотези, які не можна підтвердити або спростувати. Вони так і залишаються гіпотезами (наприклад, гіпотези походження мови).

Аналіз і синтез як загальнонаукові методи являють собою діалектичну єдність протилежностей: перший передбачає поділ цілого на частини і опис кожної з частин та зв'язків між ними, другий – поєднання частин у цілісну систему.

Таксономія є методом класифікації досліджуваних явищ і припускає їхню диференціацію відповідно до принципів двозначної логіки за єдиним параметром. Класифікації можуть мати різні параметри. У такому випадку кожне явище матиме набір ознак, кожна з яких представлятиме лише один параметр диференціації, тобто скільки висунуто параметрів класифікації, стільки встановлюється диференціацій. При використанні принципів багатозначної логіки кожна одиниця таксономії має комбінаторику ознак, причому деякі з них визначаються за принципом наявності / відсутності.

Порівняння використовується з метою виявлення загальних і специфічних ознак двох явищ, розбіжностей між ними за певними параметрами. У мовознавстві порівняння є базою порівняльно-історичного, порівняльного, типологічного та інших методів.

Формалізація як загальнонауковий метод ґрунтується на встановленні відповідності досліджуваних об'єктів, їхніх ознак і відношень до постійних, добре доступних для огляду і можливих для ототожнення матеріальних конструкцій, які дозволяють виявити і зафіксувати сутнісні особливості предмета вивчення. Формалізація передбачає використання метамови опису. У мовознавстві такими метамовами є мови формальної логіки, семантичного запису, різні типи lingua mentalis (мови думки), зокрема, пропозиціональне, фреймове, ситуаційне подання і т.д.

Ідеалізація є процесом розумового створення ідеалізованих об'єктів на підставі припущення тотожності ідеальної моделі і реального зразка, формулювання ідеалізованих припущень. Вона являє собою результат пізна-вальної діяльності дійсного стану речей і є важливим засобом створення наукових теорій. У мовознавстві результатом ідеалізації є моделі семіозиса, знака, комунікативної ситуації, дискурсу тощо.

Експеримент передбачає емпіричну перевірку гіпотези на практиці, вирішення проблеми дослідним шляхом на підставі анкетування, опиту-

вання, вимірювання, статистичного аналізу та обробки отриманих результатів. У мовознавстві широко застосовується асоціативний експеримент, різні методики експериментальних фонетичних, психолінгвістичних, нейролінгвістичних досліджень.

Фальсифікація полягає в умисному висуненні неправдивої гіпотези для встановлення об'єктивних закономірностей або взагалі доведення помилковості гіпотези шляхом її емпіричної перевірки, адже будь-яка гіпотеза повинна припускати можливість свого спростування.

Моделювання являє собою сукупність способів ідеалізації і абстрагування, аналізу і синтезу з метою розумового і формалізованого представлення об'єктів (оригіналів) дослідження і вивчення на підставі цього представлення відповідних явищ, ознак, процесів, зв'язків, передбачення і з'ясування закономірностей існування і функціонування об'єктів. Моделювання застосовується при неможливості або ускладненні дослідження оригіналів в природному середовищі для визначення їхніх характеристик, оптимізації управління та користування об'єктами для перевірки гіпотез тощо. Моделлю є будь-яка система, що має розумове уявлення, матеріально реалізується і має здатність заміщати об'єкт дослідження так, щоб його вивчення надало нову інформацію про цей об'єкт.

Лінгвістичні методи використовуються з метою вивчення мови, продуктів мови і мовної діяльності, є спеціальними і мають власну специфіку, на відміну від методів інших наук.

Систематизація загальних лінгвістичних методів у мовознавстві здійснювалася відповідно до різних напрямків або аспектів мовознавства. Так, Б. Головін розрізняє вісім загальних методів лінгвістичного дослідження: описовий, порівняльно-історичний, історичний, порівняльний, структурний, стилістичний, кількісний, автоматичного аналізу. В. Кодухов називає основними тільки три методи: описовий, порівняльний та нормативно-стилістичний, вказавши на можливість варіювання залежно від особливостей досліджуваних явищ і конкретних завдань. Мовознавець Ю. Карпенко розподіляє методи мовознавства за метою дослідження мови на описові (експериментальний, дистрибутивний, статистичний) і рекон-

структивні (порівняльно-історичний, метод внутрішньої реконструкції), за способом дослідження – на синхронічні та діахронічні; за шляхом досягнення мети – на індуктивні і дедуктивні. М. Кочерган розрізняє описовий, порівняльно-історичний, метод лінгвістичної географії, порівняльний, структурний, соціолінгвістичний, психолінгвістичний та математичний методи.

Загальними лінгвістичними методами можна назвати *парадигмальні*: порівняльно-історичний, структурний, функціональний і конструктивний. Кожен з цих методів має розгалужену систему прийомів дослідження і опису мови.

До того ж окремими методами мовознавства (*міжпарадигмальними*) можна вважати такі, що засновані на поєднанні методологічних основ парадигмальних методів, проте виділяються за специфікою процедур або з загальним підходом до аналізу мовних явищ. Так, типологічний та зіставний методи перенесли процедури порівняння на зіставлення різних мов в розрізі синхронії, властивої структурним методам, тобто виникли на межі двох наукових парадигм. Описовий метод можно віднести як до структурних, так і до функціональних методів за принципом синхронічного опису.

Окремо слід розглядати *методи маргінальних галузей*, як психолінгвістичні, соціолінгвістичні, етнолінгвістичні.

Чимало методів і методик сучасної лінгвістики є *комплексними* і комбінують процедури багатьох базових методів і методик маргінальних галузей. Наприклад, концептуальний аналіз застосовує етимологічний, компонентний, дистрибутивний, валентностний, контекстологічний, текстово-інтерпретаційний аналіз, процедури конструктивного методу, методика асоціативного експерименту, опитування, анкетування, спостереження і т.п.

Отже, всі методи сучасного мовознавства можна диференціювати на парадигмальні, міжпарадигмальні, маргінальні і комплексні або комбіновані.

Порівняльно-історичний метод являє собою сукупність операцій реконструкції походження мов від прамови, встановлення еволюційних змін і закономірностей розвитку споріднених мов шляхом їхнього порівняння на різних етапах формування. Він застосовує процедури встановлення спорідненості мов, порівняння мовних фактів у споріднених мовах, реконструкції праформ відповідної прамови, проявлення законів мовного розвитку, з огляду на закономірний еволюційний характер мовних змін.

Прийомами порівняльно-історичного методу вважаються *внутрішня і зовнішня реконструкція*. Перша спрямована на:

– відновлення попереднього стану мови, звукових, морфологічних форм, лексем і т.д.;

– інвентаризацію варіантів різних підсистем мови;

– класифікацію за давністю виникнення на підставі показників тільки однієї мови, взятої в її синхронному стані, тобто процедури внутрішньої реконструкції засновані на принципах системності та синхронії.

Кінцевою метою внутрішньої реконструкції є відтворення історії розвитку певної мови.

Їй протистоїть зовнішня реконструкція, яка сприяє відновленню попереднього стану мови шляхом зіставлення генетично споріднених мов. У рамках цього методу застосовуються також методики відносної хронології, що полягає у встановленні хронологічної послідовності зміни мовних явищ; а також лінгвогеографічна (ареальна) методика, що може надати пояснення історичних фактів шляхом картографування поширення певних мовних явищ у відповідних регіонах. Методика лінгвогеографії передбачає визначення меж та регіону функціонування діалектів і мов та інтерпретації просторового поширення мовних явищ на підставі позначення їх на географічній карті.

Структурний метод слугує для пізнання внутрішньої організації мови як системи з притаманними їй інваріантними елементами в мові, які співвідносяться з регламентованими конкретними реалізаціями; відношеннями між елементами, які впорядковують елементи в ієрархії мовних рівнів.

Найбільш дієвими і поширеними в сучасних дослідженнях є структурні методики:

1) опозиційного аналізу;

2) дистрибутивного аналізу, спрямованого на встановлення характеристик і функціональних властивостей мовної одиниці на підставі її оточення (дистрибуції), представленої одиницями того ж рівня мовної системи;

3) трансформаційного аналізу;

4) аналізу безпосередніх складових;

5) компонентного аналізу, метою якого є встановлення структури значення слова як певним чином організованої сукупності елементарних смислових одиниць – сем (семантичних множників);

6) комутації;

7) ланцюгового аналізу, що передбачає дослідження лінійних синтаксичних структур, представлених ядерними конструкціями і право- і лівобічними розповсюджувачами – багатокомпонентними ланцюжками, які членують на інші елементарні одиниці.

Функціональний метод передбачає дослідження мови в дії, в процесі функціонування з огляду на цілеспрямовану природу мовних одиниць і явищ. Цей метод застосовує методики моделювання функціонально-семантичних полів, контекстуально-інтерпретаційного аналізу тексту, прагматичного і конверсаційного аналізу, аналізу дискурсу, діалогічної інтерпретації тексту і т.п.

Конструктивний метод також є загальним методом лінгвістики, передбачає побудову та конструювання об’єкта дослідження у вигляді спрощеної, гіпотетичної абстрактної схеми. Він покладений в основу генеративного методу моделювання глибинних синтаксичних структур, встановлення трансформацій і обмежень на них, принципів і параметрів бази мовних знань Н. Хомського і його послідовників, моделювання позиційних схем сучасного семантичного синтаксису на підставі граматики залежностей Л. Теньєр, рольової граматики Ч. Філлмора, речень

Б. Рассела, функціонально-істиннісної граматики Р. Монтегю і т.д., концепцій геологічного аналізу мови і т.п.

Метою ***типологічного методу*** як міжпарадигмального є диференціація мов світу залежно від їхніх структурних, граматичних і функціональних характеристик безвідносно до генетичного спорідненння. Головним принципом типологічного методу є порівняння як встановлення подібності і відмінностей між мовами. Базовим поняттям такої диференціації є тип мови, що передбачає сукупність певних семантико-граматичних характеристик при домінуванні найбільш загальної та імплікації інших. Шляхом типологічного методу здійснюється аналіз синхронного стану мов незалежно від їхньої генетичної спорідненості або віддаленості.

Типологічна класифікація мов може бути системно-цілісною і фрагментарною. Перша ґрунтується на домінантній типологічній ознаці, друга вибирає параметром типологізації конкретну ознаку одиниць певного мовного рівня (фонологічного, деривaційного, синтаксичного і т.д.).

Зіставний метод спрямований на виявлення загальних і специфічних рис порівняних мов на всіх рівнях і в мовленні, тексті за принципом синхронії. Зіставний метод застосовується з різною метою:

– поглиблення знань про одну з мов на тлі іншої;

– виявлення особливостей взаємодії в процесі оволодіння іншою мовою;

– встановлення характерних рис сімей і груп мов;

– прогнозування інтерференції мов;

– визначення оптимальних засобів оволодіння іншою мовою і т.п.

Об’єктом зіставного аналізу можуть бути фонеми, склади, лексичні одиниці, граматичні категорії, номінативні структури, синтаксичні конструкції, тексти.

Описовий метод являє собою сукупність процедур інвентаризації, таксономії та інтерпретації досліджуваних мовних явищ на певному етапі розвитку мови (в синхронії). Мовознавці представляють описовий метод як послідовність етапів:

1) виділення одиниць аналізу (інвентаризації);

2) членування виділених одиниць на менші складові частини (сегментації);

3) класифікації одиниць (таксономії);

4) виявлення ознак груп таксономії (інтерпретації).

Описовий метод застосовує прийоми внутрішньої і зовнішньої інтерпретації.

У зв'язку зі становленням маргінальних галузей мовознавства реєстр його методів значно поповнився за рахунок психолінгвістичних, соціолінгвістичних та етнолінгвістичних методик, зокрема, *психолінгвістичного експерименту, безпосереднього і включеного спостереження, опитування, прослуховування звукозаписів, анкетування, тестування, інтерв'ювання із залученням лінгвостатистичних методів (лінгвостатистичного експерименту, факторного, кореляційного аналізу)*.

Загалом процедурне поле сучасної лінгвістичної методології можна кваліфікувати як комбіноване, комплексне. Розвинений методологічний інструментарій сучасної лінгвістики сприяє більшій об'єктивності і достовірності досліджень такого багатовимірного і динамічного феномена, яким є природна мова [4, с. 48–64].

Завдання

1. Описати зв'язки лінгвістики з іншими науковими дисциплінами.
2. Навести приклади застосування лінгвістичних методів.

Література

1. Соснина Е. П. Введение в прикладную лингвистику / Е. П. Соснина. – Ульяновск : УлГТУ, 2012. – 110 с.

2. Комарова З. И. Методология, метод, методика и технология научных исследований в лингвистике / З. И. Комарова. – Екатеринбург : Изд-во УрФУ, 2012. – 818 с.

3. Зубов А. В. Информационные технологии в лингвистике / А. В. Зубов, И. И. Зубова. – Москва : Академия, 2004. – 208 с.

4. Селіванова О. О. Сучасна лінгвістика: напрями та проблеми / О. О. Селіванова. – Полтава : Довкілля-К, 2008. – 712 с.

5. Арнольд И. В. Основы научных исследований в лингвистике / И. В. Арнольд. – Москва : Высш. шк., 1991. – 140 с.

6. Баранов А. Н. Введение в прикладную лингвистику / А. Н. Баранов. – Москва : ЛКИ, 2007. – 360 с.

7. Schmitt N. An Introduction to Applied Linguistics / N. Schmitt, M. P. H. Rodgers. – London: Routledge, 2019. – 392 p.

8. Davies A. An Introduction to Applied Linguistics. From Practice to Theory / A. Davies, K. Mitchell. – Edinburgh University Press, 2007. – 199 p.

9. Ojeda A. E. A Computational Introduction to Linguistics / A. E. Ojeda. – Centre for the Study of Language & Information, 2015. – 400 p.

10. Akmajian A. Linguistics: An Introduction to Language and Communication / A. Akmajian, A. K. Farmer, L. Bickmore, R. A. Demers, R. M. Harnish. – The MIT Press, 2017. – 608 p.

РОЗДІЛ 2
МОДЕЛЮВАННЯ ПРИРОДНОЇ МОВИ ЗАСОБАМИ СТРУКТУРНОЇ ТА МАТЕМАТИЧНОЇ ЛІНГВІСТИКИ

2.1. Поняття лінгвістичної моделі.

2.2. Формальні мови та граматики.

2.3. Особливості моделювання природної мови.

2.1. Поняття лінгвістичної моделі

Математичну лінгвістику визначають як математичну дисципліну, об'єктом дослідження якої є природні мови.

Під ***математичною лінгвістикою*** розуміють не тільки теорію граматик, а й взагалі використання понять, загальнологічних в своїй основі (множина, відношення, упорядкована послідовність, опозиція і т.п.) – понять, які відіграють істотну роль в структурній лінгвістиці, де вони становлять саме основу всього наукового апарату.

Сучасну ***структурну лінгвістику*** можна охарактеризувати як галузь лінгвістики, ядром якої є деяка область фактів, які підлягають компетенції як математики, так і лінгвістики, яка поширюється далі на всі питання, пов'язані з взаємозв'язком інших аспектів структури мови з цим ядром.

Сучасна структурна лінгвістика включає в себе як процедури виявлення, так і апарат породжувальних граматик, маючи за мету об'єднання всіх засобів точного опису мови.

Сукупність уявлень, спільних для структурної і математичної лінгвістики, можна коротко охарактеризувати як область моделювання мови [1, с. 16–29].

Необхідність у моделюванні виникає у всіх тих наукових галузях, де об'єкт науки недоступний безпосередньому спостереженню.

Найбільш важливі властивості моделей, у тому числі лінгвістичних:

1. Моделювати можна тільки такі явища, істотні властивості яких вичерпуються їхніми структурними (функціональними) характеристиками.

До явищ, суттєві ознаки яких зводяться до функціональних або структурних властивостей, належить мова.

2. Модель завжди є деякою ідеалізацією об'єкта. Щоб зрозуміти складні явища, необхідно почати з вивчення найпростіших і загальних випадків і від них просуватися до більш складних і спеціальних випадків.

Наприклад, у лінгвістиці принцип «підстановки» відсутніх членів речення являє собою випадок ідеалізації лінгвістичних об'єктів. Принцип підстановки часто використовується в синтаксисі при описі деяких типів речень: одночленні (безособові) речення можна розглядати як продукт скорочення нормальних двочленних речень (Морозить. – Мороз морозить.).

3. Зазвичай модель оперує не поняттями про реальні об'єкти, а конструктами, тобто поняттями про ідеальні об'єкти, які не виводяться безпосередньо і однозначно з дослідних даних, але побудованими на підставі деяких загальних гіпотез. Будь-яка модель є конструкцією, логічно виведеною з гіпотез за допомогою певного математичного апарату.

4. Будь-яка модель, у тому числі лінгвістична, повинна бути формальною. Модель вважається формальною, якщо в ній в явному вигляді і однозначно задані вихідні об'єкти, що зв'язують їхні твердження і правила поводження з ними.

Формальна модель зв'язується з дослідними даними за допомогою тієї чи іншої інтерпретації. Дати інтерпретацію моделі – означає вказати правила, імовірнісні або строгі, постановки об'єктів деякої предметної області.

5. Будь-яка інтерпретована модель, у тому числі лінгвістична, повинна мати властивість експланарності, або пояснювальної сили. Вважається, що модель володіє цією властивістю, якщо вона:

1) пояснює факти або дані спеціально поставлених експериментів, які незрозумілі з точки зору старої теорії;

2) передбачає невідому раніше, але принципово можливу поведінку об'єкта, яка пізніше підтверджується даними спостереження або нових експериментів.

Отже, побудова моделі передбачає:

1) фіксування фактів, які потребують пояснення;

2) висування гіпотез для пояснення фактів;

3) реалізацію гіпотез у вигляді моделей, які не тільки пояснюють вихідні факти, а й пророкують нові факти, які ще не спостерігалися;

4) експериментальну перевірку моделі.

Основні типи лінгвістичних моделей, що відрізняються один від одного за характером розглянутого в них об'єкта:

1) моделі, в яких як об'єкт розглядаються процедури, що ведуть до виявлення того чи іншого мовного явища. Ці моделі імітують дослідницьку діяльність лінгвіста;

2) моделі, в яких як об'єкт виступають конкретні мовні процеси і явища. Ці моделі імітують мовну діяльність людини.

Дослідні моделі можна поділити на три класи залежно від того, яка інформація використовується в них як вихідна. У моделях першого класу як вихідна інформація використовується тільки текст, і всі відомості про систему, тобто мову, що породжує цей текст, витягуються виключно з текстових даних. Це класичні дешифрувальні моделі.

У моделях другого класу вважається заданим не лише текст, а й множина правильних фраз конкретної мови. Це означає, що при розробці моделі лінгвіст вдається до допомоги інформанта, який з приводу кожної пропонованої йому фрази повинен говорити, правильна вона чи ні. Інформантом може бути і сам лінгвіст, якщо він досконало володіє мовою, що вивчається.

У моделях третього класу вважаються заданими не тільки текст і множина правильних фраз, але і множина семантичних інваріантів. Це означає, що інформант повинен визначати не тільки правильність кожної пропонованої йому фрази, а й говорити про будь-які дві фрази, означають вони одне і те ж чи ні.

Моделі конкретних мовних процесів і явищ або ***моделі мовної діяльності людини*** діляться на моделі аналізу, моделі синтезу і породжувальні моделі. *Моделлю аналізу* називається скінченна кількість правил, здатних проаналізувати нескінченне число правил, здатних проаналізувати нескінченне число речень конкретної мови. Синтаксичні аналітичні моделі отримують на вході текст, а на виході видають для кожного речення запис його синтаксичної структури. Семантичні аналітичні моделі отримують на вході той же матеріал, а на виході видають смисловий запис кожного речення спеціальною семантичною мовою.

Моделлю синтезу називається скінченне число правил, здатних побудувати нескінченно велике число правильних речень. Синтаксичні синтетичні моделі використовують в якості вихідної інформації запис синтаксичної структури речень, а на виході видають правильні речення конкретної мови. Семантичні синтетичні моделі отримують на вході смисловий запис деякого речення на спеціальній семантичній мові і видають на виході множину речень природної мови, синонімічних цьому реченню.

Породжувальною моделлю називається пристрій, що містить алфавіт символів і скінченне число правил творення і перетворення виразів з елементів цього алфавіту, здатне побудувати множину правильних речень конкретної мови і приписати кожній з них деяку структурну характеристику [2, с. 78–112].

Відношення між розглянутими типами лінгвістичних моделей зображені в таблиці 2.1.

Таблиця 2.1 – Типи лінгвістичних моделей

Ознаки / Тип моделі	Що відомо лінгвісту	Вихідна інформація	Кінцева інформація	Мета
Дослідні	Текст (та множина правильних фраз)	Текст	Граматика та словник	Змоделювати діяльність лінгвіста

Продовження табл. 2.1

Ознаки / Тип моделі	Що відомо лінгвісту	Вихідна інформація	Кінцева інформація	Мета
Аналітичні	Граматика та словник	Текст	Синтаксична структура речень	Змоделювати розуміння тексту
Синтетичні	Граматика та словник	Синтаксична структура речень	Текст	Змоделювати творення тексту
Породжувальні	Граматика та словник	Алфавіт символів і правила творення та перетворення фраз	Множина правильних фраз та їхня синтаксична структура	Змоделювати вміння відрізняти правильне від неправильного в мові

2.2. Формальні мови та граматики

Формальні граматики – спосіб опису формальної мови (множина скінченних слів (рядків, ланцюжків) над скінченним алфавітом).

Породжувальні граматики. Якщо мова L складається з невеликого числа ланцюжків, то найочевидніший спосіб опису мови – скласти список всіх ланцюжків з L.

Однак багато мов, наприклад, мови програмування неможливо або небажано ставити вичерпним переліком ланцюжків, що до них входять. Тому, як правило, використовуються інші способи визначення мови, які дозволяють опису мови бути доступним для огляду (скінченним), хоча мова, що описується, може бути і нескінченною.

У 1956 р. Н. Хомський запропонував ієрархію формальних мов, породжуваних граматиками, за видом їхніх правил. Він виділив чотири типи граматик (рис. 2.1): регулярні (Р), контекстно-вільні (КВ), контекстно-залежні (КЗ) і необмежені (Н) (з фразовою структурою).

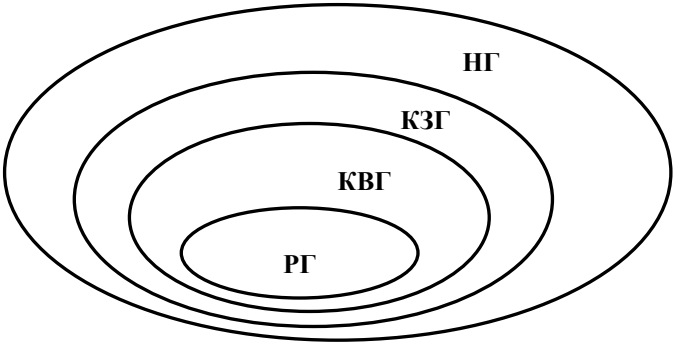

Рисунок 2.1 – Ієрархія Хомського формальних мов

Кожен наступний тип граматик формальних мов є розширенням попереднього (під «розширенням» тут розуміється збільшення складності мови).

Таким чином, породжувальні граматики (генеративні граматики) задають правила, за допомогою яких можна побудувати будь-який ланцюжок мови. *Генеративна граматика* – лінгвістична теорія, яка розглядає граматику як систему правил, яка генерується в точності з тих комбінацій слів, які складають граматичні речення цієї мови. Одна з переваг визначення мови за допомогою породжувальних граматик полягає в тому, що граматика надає ланцюжкам мови певну структуру, яка в більшості випадків може відображати зміст речення.

Розпізнавальні граматики (аналітичні граматики) дозволяють за словом визначити, входить воно в мову чи ні. Цей метод визначення мови пов'язаний зі способом задання множини за допомогою характеристичної властивості (предиката) і полягає у використанні часткового алгоритму, який для довільного вхідного ланцюжка зупиниться і відповість «так»

після кінцевого числа кроків, якщо цей ланцюжок належить мові. Схематизовані пристрої, що використовуються для представлення таких алгоритмів, називаються розпізнавачами. Прикладами розпізнавачів є скінченні автомати, автомати з магазинною пам'яттю і машини Тьюринга [3, с. 53–54].

2.3. Особливості моделювання природної мови

Складність моделювання пов'язана з тим, що природна мова – велика відкрита багаторівнева система знаків, що виникла для обміну інформацією в процесі практичної діяльності людини, і постійно змінюється в зв'язку з цією діяльністю.

Одним із наслідків тривалого історичного розвитку природної мови є нестандартна сполучуваність (синтактика) одиниць на кожному рівні мови. На відміну від штучних формальних мов (мов логіки, мов програмування), в яких сполучуваність знаків диктується їхньою семантикою та може бути зафіксована синтаксично (граматично), в природних мовах поєднання слів у реченнях лише частково може бути описано законами граматики.

Однією із найбільших складнощів при обробці природно-мовних текстів є *неоднозначність (багатозначність)*, що виявляється на всіх рівнях природної мови, що виражається в явищах полісемії, омонімії, синонімії.

Полісемія – наявність у однієї одиниці мови кількох пов'язаних між собою значень. *Синонімія* – повний або частковий збіг значень різних одиниць. *Омонімія* – збіг за формою двох різних за змістом одиниць (на відміну від полісемії немає смислового зв'язку між одиницями, що збігаються за формою).

Складність формального опису природної мови та її обробки веде до розбиття цього процесу на окремі етапи, відповідні рівням мови.

Існує кілька способів розбиття тексту на одиниці, що належать до різних рівнів:

– рівень речень (висловлювань) – синтаксичний рівень;

– рівень слів (словоформ – слів у певній граматичній формі) – морфологічний рівень;

– рівень фонем (окремих звуків, за допомогою яких формуються і розрізняються слова) – фонологічний рівень.

Більшість сучасних лінгвістичних процесорів належать до модульного типу, в якому кожному рівню / етапу аналізу або синтезу тексту відповідає окремий модуль процесора.

У разі аналізу тексту окремі модулі виконують:

– графематичний аналіз (сегментацію), тобто виділення в тексті речень і словоформ, точніше токенів (тому що в тексті можуть бути не тільки слова) – перехід від символів до слів;

– морфологічний аналіз – перехід від словоформ до їхньої леми (словникових форм лексем) або основ (ядерних частин слова, за винятком словозмінних морфем);

– синтаксичний аналіз – виявлення синтаксичних зв'язків слів і граматичної структури речень;

– семантичний і прагматичний аналіз, при якому визначаються сенс фраз і відповідна реакція системи.

Модулі ***морфологічного аналізу*** розрізняються за методом аналізу – з опорою на словник словоформ мови або на словник основ та безсловниковий метод. При морфологічному синтезі вихідними даними є лексема і конкретні морфологічні характеристики словоформи цієї лексеми, можливий і запит на синтез усіх форм заданої лексеми (так званої парадигми слова).

Прикладами модулів морфологічного аналізу є Pymorphy, TreeTagger, Snowball та ін.

Наприклад, TreeTagger – морфологічний аналізатор, розроблений Хелмут Шмідом в інституті комп'ютерної лінгвістики університету Штутгарта. TreeTagger оперує деревами прийняття рішень і застосовується в

задачах обробки російської, англійської, німецької, французької та інших мов (рис. 2.2).

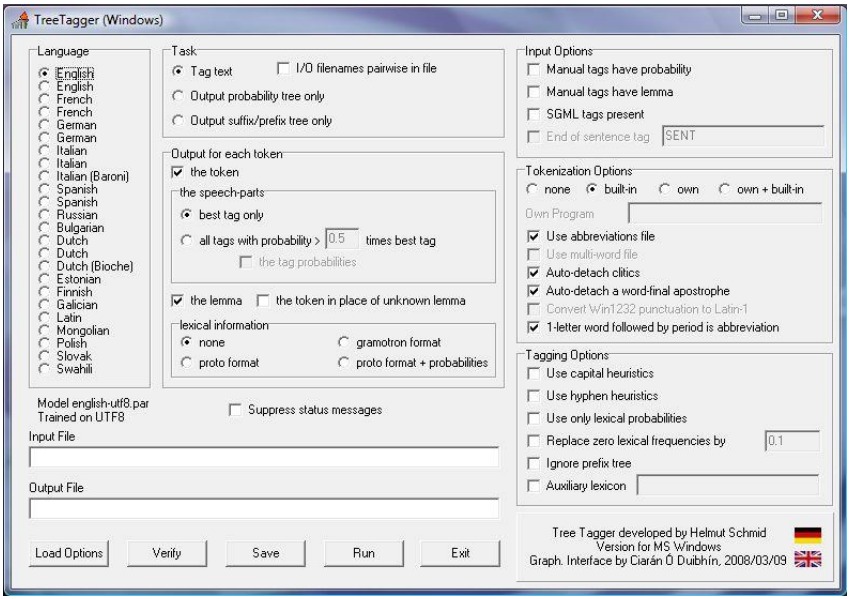

Рисунок 2.2 – Морфологічний аналізатор TreeTagger

Для реалізації ***синтаксичного етапу*** запропоновано велику кількість різних ідей і методів, що відрізняються способом опису синтаксису мови, способом використання цієї інформації при аналізі або синтезі речень, а також способом представлення синтаксичної структури речення. Можна виділити три основні підходи:

– генеративний підхід, висхідний до ідей породжувальних граматик Н. Хомського;

– підхід, висхідний до ідей І. Мельчука і поданий у лінгвістичній моделі «Смисл⇔Текст»;

– підхід, в рамках якого робляться ті чи інші спроби подолати обмеження перших двох підходів, зокрема, теорія синтаксичних груп.

У рамках *генеративного підходу* синтаксичний аналіз проводиться, як правило, на основі формальної контекстно-вільної граматики, яка описує фразову структуру речення, або ж на основі деякого розширення контекстно-вільної граматики. Ці граматики виходять з послідовного лінійного членування речення на фрази (різні словосполучення) і тому відображають одночасно як його синтаксичну, так і лінійну структури. Отримана в результаті ієрархічна синтаксична структура речення описується деревом складових, у листі якого знаходяться слова речення, піддерева відповідають синтаксичним конструкціям (фразам), що входять в речення, а дуги виражають відношення вкладення конструкцій.

У рамках другого підходу для подання синтаксичної структури речення використовуються *дерева залежностей*. У вузлах дерева розташовані слова речення (у корені – слово-предикат, зазвичай дієслово-присудок), а кожна дуга дерева, що зв'язує пару вузлів та інтерпретується як синтаксичний підрядний зв'язок між ними, причому напрямок зв'язку відповідає напрямку дуги. Оскільки при цьому синтаксичні зв'язки слів і порядок слів у реченні відокремлені, то на основі дерев підпорядкування можуть бути описані розірвані і непроективні конструкції, що досить часто виникають у мовах з вільним порядком слів.

Дерева складових більше підходять для опису мов з жорстким порядком слів, подання за їх допомогою розірваних і непроективних конструкцій вимагає розширення граматичного формалізму. Загальні труднощі для обох підходів – представлення однорідних членів речення.

Синтаксичні моделі в усіх описаних підходах намагаються врахувати обмеження, що накладаються на поєднання мовних одиниць у мовленні, при цьому так чи інакше використовується поняття валентності. *Валентність* – здатність слова або іншої одиниці мови приєднувати інші одиниці певним синтаксичним способом; *актант* – слово або синтаксична конструкція, що заповнює цю валентність. У рамках *генеративного підходу* валентності слів описуються переважно у вигляді спеціальних фреймів, а в рамках підходу, заснованого на деревах залежностей, – як моделі управління.

Модулі синтаксичного аналізу в обох розглянутих підходах спираються на граматики природної мови. Загальна кількість правил граматики може бути від декількох десятків до декількох сотень, залежно від словника, що використовується: чим більше інформації подано у словнику, тим коротше може бути граматика, і навпаки. Так, у моделі «Смисл⇔Текст» наголос робиться на словник, а не на граматику; у словнику зберігається інформація, яка належить до різних рівнів мови, зокрема, про моделі управління слів і нестандартну сполучуваність слів.

Етап *семантичного аналізу* тексту найменш опрацьований. Для локального семантичного аналізу, тобто аналізу речень були запропоновані так звані відмінкові граматики і семантичні відмінки (валентності), на базі яких семантика речень описується через зв'язки головного слова з його семантичними актантами, тобто через семантичні відмінки.

Для подання семантики всього тексту зазвичай використовуються два формалізми:

1) формули обчислення предикатів, що виражають властивості, стан, процеси, дії і відношення;

2) семантичні мережі – розмічені графи, в яких вершини відповідають поняттям, а дуги – відношенням між ними.

Крім багаторівневості системи мови складність її моделювання пов'язана зі змінами, що постійно в ній відбуваються. Зміни стосуються не тільки словникового запасу мови, але також синтаксису, морфології та фонетики [4, с. 14–21].

Завдання

Навести приклади сучасних лінгвістичних аналізаторів для кожного рівня природної мови (графематичного, морфологічного, синтаксичного та семантичного). Описати підходи і методи, які в них використовуються.

Література

1. Ревзин И. И. Современная структурная лингвистика. Проблемы и методы / И. И. Ревзин. – Москва : Изд-во Наука, 1977. – 263 с.

2. Апресян Ю. Д. Идеи и методы современной структурной лингвистики / Ю. Д. Апресян. – Москва : Изд-во Просвещение, 1966. – 302 с.

3. Батура Т. В. Математическая лингвистика и автоматическая обработка текстов / Т. В. Батура. – Новосибирск : РИЦ НГУ, 2016. – 166 с.

4. Автоматическая обработка текстов на естественном языке и анализ данных / Е. И. Большакова, К. В. Воронцов, Н. Э. Ефремова, Э. С. Клышинский, Н. В. Лукашевич, А. С. Сапин. – Москва : Изд-во НИУ ВШЭ, 2017. – 269 с.

5. Хомский И. Введение в формальный анализ естественных языков / И. Хомский, Дж. Миллер. – Москва : Едиториал, 2003. – 64 с.

6. Обробка тексту [Електронний ресурс]. – Режим доступу : https://nlpub.ru/Обработка_текста.

7. Schmitt N. An Introduction to Applied Linguistics / N. Schmitt, M. P. H. Rodgers. – London: Routledge, 2019. – 392 p.

8. Kurdi M. Z. Natural Language Processing and Computational Linguistics: Speech, Morphology and Syntax / M. Z. Kurdi. – Wiley, 2016. – 296 p.

РОЗДІЛ 3
АКТУАЛЬНІ ПРОБЛЕМИ КОМП'ЮТЕРНОЇ ЛІНГВІСТИКИ

3.1. Прикладні завдання комп'ютерної лінгвістики.
3.2. Машинний переклад.
3.3. Автоматична генерація текстів.
3.4. Створення діалогових систем.
3.5. Розпізнавання та синтез мовлення.
3.6. Класифікація та кластеризація.
3.7. Витяг інформації.
3.8. Аналіз тональності текстів.

3.1. Прикладні завдання комп'ютерної лінгвістики

Комп'ютерна лінгвістика – напрям прикладної лінгвістики, що вивчає лінгвістичні основи інформатики та всі аспекти зв'язку мови і мислення, моделювання мови та мислення в комп'ютерному середовищі за допомогою комп'ютерних програм. Цей напрям застосовують у такій області:

– оптимізації комунікації на основі лінгвістичних знань;

– створення природно-мовного інтерфейсу і технологій розуміння мови для спілкування людини з машиною [1, с. 27].

Комп'ютерна лінгвістика – міждисциплінарна галузь, яка виникла на стику таких наук, як лінгвістика, математика, інформатика, штучний інтелект.

Витоки комп'ютерної лінгвістики досягають дослідження американського лінгвіста Н. Хомського з формалізації структури природної мови, до перших експериментів з машинного перекладу, виконаних програмістами і математиками, а також до розроблених в області штучного інтелекту перших програм розуміння природної мови.

Прикладні завдання, які вирішуються інструментами комп'ютерної лінгвістики:

Машинний переклад (Machine Translation) – напрям комп'ютерної лінгвістики, разом з яким виникла і розвивалася ця галузь.

Ще одне прикладне завдання комп'ютерної лінгвістики – це ***інформаційний пошук (Information Retrieval)*** і пов'язані з ним завдання індексування, реферування, класифікації та рубрикації документів.

Реферування тексту (Summarization) – скорочення його обсягу і отримання короткого викладу його змісту – реферату, що робить більш швидким пошук у колекціях документів.

Близьке до реферування завдання – ***анотування тексту*** документа, тобто складання його анотації. У простій формі анотація являє собою перелік основних (ключових) елементів тексту, для виділення яких використовуються статистичні та лінгвістичні критерії.

При обробці великих колекцій документів актуальні завдання ***класифікації (Categorization)*** і ***кластеризації текстів (Text Clustering)***. Класифікація означає віднесення кожного документа до певного класу із заздалегідь відомими параметрами, а кластеризація – розбиття множини документів на кластери, тобто підмножини тематично близьких документів.

Актуальна прикладна задача, яка часто належить до напрямку Text Mining – це ***витяг інформації з текстів (Information Extraction)***. При вирішенні цього завдання здійснюється виділення в природно-мовному тексті певних об'єктів – іменованих сутностей (імен персоналій, географічних назв, назв фірм та ін.), їхніх відношень і пов'язаних з ними подій.

Виділення думок (Opinion Mining) і ***аналіз тональності текстів (Sentiment Analysis)***: у першій задачі відбувається пошук (у блогах, форумах, інтернет-магазинах та ін.) думок користувачів про товари та інші об'єкти, а також проводиться аналіз цих думок. Друге завдання близьке до класичної задачі контент-аналізу текстів масової комунікації, у ньому оцінюється загальна тональність висловлювань і тексту в цілому.

Ще одна прикладна задача, яка виникла більше 50 років тому і розвиток якої стимулювало появу мережі Інтернет, – це ***підтримка природно-мовного діалогу***.

Зовсім інший прикладний напрямок, який розвивається, – ***автоматизація підготовки та редагування текстів***. Одними з перших досягнень у цьому напрямку були програми автоматичного визначення переносів слів та програми орфографічної перевірки тексту (правопис, або автокоректор).

Наступний прикладний напрямок, який варто згадати, – це ***автоматична генерація текстів***.

Напрямками, які активно розвиваються, є ***розпізнавання і синтез мовлення***.

Комп'ютерна лінгвістика демонструє результати в різних прикладних напрямках з автоматичної обробки та аналізу текстів. У більшості додатків використовуються прості і редуковані моделі, які дають прийнятні результати; нерідко якість результатів досягає експертного рівня – зазвичай там, де думки експертів можуть розходитися.

Подальший прогрес в області комп'ютерної лінгвістики пов'язаний як з більш точним урахуванням лінгвістичних особливостей текстів на різних етапах його обробки та застосуванням більш детальних лінгвістичних моделей, так і з розвитком методів машинного навчання і пошуком більш ефективних методів та їхніх комбінацій для кожної прикладної задачі [2, с. 7–13].

3.2. Машинний переклад

Сьогодні існують три різних підходи до машинного перекладу:

1. Переклад на основі правил (rule-based machine translation): необхідні якомога детальніший словник і дуже докладна граматика обох мов. Усі ранні системи перекладів були засновані на правилах.

2. Статистичний машинний переклад (statistical machine translation) обходиться без словника і граматики, він працює тільки на основі методів машинного навчання. Програмі потрібно дуже багато пар «речення + його переклад». Вона запам'ятає, як перекладалися двійки або трійки слів, в якому оточенні вони зустрічалися, де в реченні знаходилися. Після цього,

отримавши для обробки нове речення, вона зможе вибрати його найбільш ймовірний переклад. При цьому вона буде враховувати тільки оточення, в якому зустрічалося те чи інше слово, і частоту різних ланцюжків слів.

3. Гібридний переклад (hybrid machine translation) – це комбінація правил і статистики. Поєднувати два ці підходи можна різними способами, тому гібридні перекладачі можуть бути різних типів. Зараз це найсучасніший підхід, майже всі розробники прагнуть використовувати гібридні технології.

При машинному перекладі текст обробляється комп'ютером без участі людини. А якщо текст перекладається людиною з використанням різних комп'ютерних технологій, то говорять про автоматизований переклад (CAT, computer-aided translation або computer-assisted translation).

Переклад на основі правил. Прийнято виділяти три різних типи перекладу на основі правил:

– трансферні системи (вихідне речення → препароване вихідне речення → речення на іншій мові);

– системи послівного перекладу (слова вихідного тексту → слова перекладу);

– інтерлінгвістичні системи (вихідний текст → опис його змісту на універсальній мові-посереднику → текст перекладу).

Трансферна система (transfer-based machine translation, transfer systems) найбільшою мірою наслідує людину-перекладача. Вона аналізує вихідне речення, перекладає його слова, з'ясовує їхні ролі, а потім за допомогою граматики збирає нове речення кінцевою мовою.

Система послівного перекладу (word-by-word machine translation, інші назви – direct systems, dictionary based machine translation) – це найпримітивніший підхід. У реченні перекладаються всі слова, а текст правиться мінімально. При цьому виходить переклад хоча і не дуже гладкий, але в більшості випадків зрозумілий. Цей підхід використовувався в ранніх програмах, він і зараз іноді застосовується, наприклад, для спрощення роботи перекладачів-людей.

Інтерлінгвістичні системи (interlingual systems) використовують мову-посередника, яка називається інтерлінгва (interlingua) – формальне подання сенсу людської мови і думок. На зорі комп'ютерної лінгвістики ідея створення такої системи була дуже популярна, але створити повноцінну універсальну допоміжну мову вченим поки не вдалося.

Прикладом системи машинного перекладу на основі правил є словник компанії PROMT. Основу системи PROMT складають словники і граматики.

Система словників у PROMT містить:

1) словники основ;

2) таблиці флексій;

3) допоміжні таблиці.

У словнику основ зберігається:

– базова форма кожного слова;

– його псевдооснова;

– його граматичні та семантичні ознаки;

– його переклад на іншу мову з підказками, який переклад в яких випадках використовувати.

У таблиці флексій зберігаються всі види словозмін.

Допоміжні таблиці створені для обробки нестандартних випадків слів. Таких таблиць кілька. Це бази префіксів і постфіксів, а також база імен та географічних назв.

Статистичний машинний переклад. Модель мови або мовна модель (language model) – спосіб обчислювати ймовірність для всіх теоретично можливих речень мови. Ці способи можуть бути різними, тому і моделі мови існують різні.

Щоб побудувати модель мови, машині потрібен корпус – великий набір текстів. При цьому бажано, щоб його тексти за стилем не надто відрізнялися від тих текстів, з якими буде працювати система. А якщо система повинна працювати з різними текстами, то і корпус повинен складатися з різних текстів.

Прикладом статистичної моделі є ланцюги Маркова, що дозволяють сильно спростити обчислення умовної ймовірності.

Гібридний переклад. Для досягнення найбільшої ефективності більшість систем машинного перекладу сьогодні використовують і правила, і статистику. Такий підхід називається гібридним.

Таким чином, переклад відбувається в два етапи:

1) переклад за правилами;

2) статистичне доопрацювання перекладеного тексту за допомогою системи, навченої на паралельному корпусі виправлень.

Сучасні системи машинного перекладу. Systran надає безкоштовний онлайн-переклад на своєму сайті http://www.systransoft.com/. Переклад розроблений для 52 мов, безкоштовна онлайн-версія доступна тільки для десяти з них.

Система OpenLogos переводить з німецької та англійської мов на французьку, італійську, іспанську та португальську; працює на основі правил і дозволяє підключати нові словники.

Linguatec засновано на правилах. У 2006 р. компанія запатентувала технологію нейронного трансферу, що працює за принципами нейронної мережі, з 2010 р. у систему стали інтегруватися гібридні технології.

Система Apertium – відкрита система, заснована на правилах. Система створювалася спочатку для близькоспоріднених мов, але зараз вона працює для різних пар. Щоб підключити нову мовну пару, потрібно додати словник і лінгвістичну інформацію про структуру необхідних мов.

Google Translate – безкоштовний перекладацький сервіс Google. Пропонує переклад для будь-якої мовної пари з 73 мов. Цей перекладач розроблений американською компанією Google. Спочатку вона пропонувала переклад на основі системи Systran. У 2004 р. керівництво компанії вирішило створити свій власний перекладач на основі статистичних методів. У 2006 р. перекладач Google діяв для пар арабська-англійська та англійська-арабська, але незабаром до нього додалися пари з російською і китайською мовами. У 2007 р. Google перевів усі наявні у нього мовні

пари на статистичні методи перекладу, відмовившись від використання перекладача Systran. Після цього в систему додавалися нові мови і опції.

Перекладач Bing розробляється компанією Microsoft з 2009 р. для її пошукового сервісу Bing. Обробляє понад 40 мов на основі статистичних методів.

У 2011 р. пошукова система Яндекс запустила сервіс «Яндекс. Перекладач», що діє на основі статистичних методів. Він дозволяв перекладати тексти і Web-сторінки з російської на англійську або українську мови і навпаки. Зараз цей сервіс надає переклад для 67 мов.

Ядром технології перекладача на основі ABBYY Compreno стала універсальна ієрархія понять і мережа відношень між ними. Ця ієрархія містить всі смисли, про які говорять і думають люди, з указанням зв'язків між цими смислами. У результаті у кожному реченні можна виділити закодований у ньому сенс [3, с. 158–189].

Сьогодні істотний інтерес представляють проекти багато-мовного перекладу з використанням проміжної мови, на якій кодується смисл фраз, що перекладаються. Сучасний напрямок – статистична трансляція, яка спирається на статистику перекладних пар слів і словосполучень. Незважаючи на багато десятиліть досліджень цього завдання, якість машинного перекладу ще далека до досконалості. Істотний прорив у цій області пов'язують з використанням машинного навчання і нейронних мереж.

3.3. Автоматична генерація текстів

Системи генерації текстів забезпечують автоматичне породження зв'язних текстів, вони покликані донести до користувача в звичній для нього формі тексту накопичені знання.

Процес генерації тексту у відповідь на питання до бази даних або бази знань проходить дві стадії. Перша визначає зміст і структуру відповіді, тобто вирішує, що треба відповідати, це стратегічний компонент (планувальник тексту).

Друга стадія – лінгвістичний (тактичний) компонент – визначає, як будувати текст відповіді, які лексичні, синтаксичні та комунікативні засоби природної мови потрібні для оформлення відповіді.

Програма, що керує ними, визначає призначення системи генерації текстів і характер бази знань, з якої береться інформація. Вона ініціює процес генерації і визначає цілі, які повинні бути досягнуті: викласти вміст бази знань у вигляді тексту, дати пояснення якоїсь функції, видати визначення якогось об'єкта і т.д.

Стратегічний компонент визначає шляхи досягнення поставлених цілей:

1) вибір інформації, яка повинна бути виражена або опущена;

2) визначення способу оформлення відповіді (перелік об'єктів, опис події);

3) структурування тексту: задання меж і порядку слідування речень;

4) вибір лексики;

5) оформлення відношень кореферентності (анафора, еліпсис);

6) вибір і порядок синтаксичних складових.

Лінгвістичний компонент породжує тексти відповідно до специфікацій планувальника. Він повинен забезпечувати граматичну правильність речень. До його компетенції входять синтаксичний і морфологічний синтез.

Приклади систем генерації тексту:

- Diogenes;
- SemSyn Spokesman;
- Gossip;
- FoG;
- LFS;
- SeoGenerator (рис. 3.1) та ін.

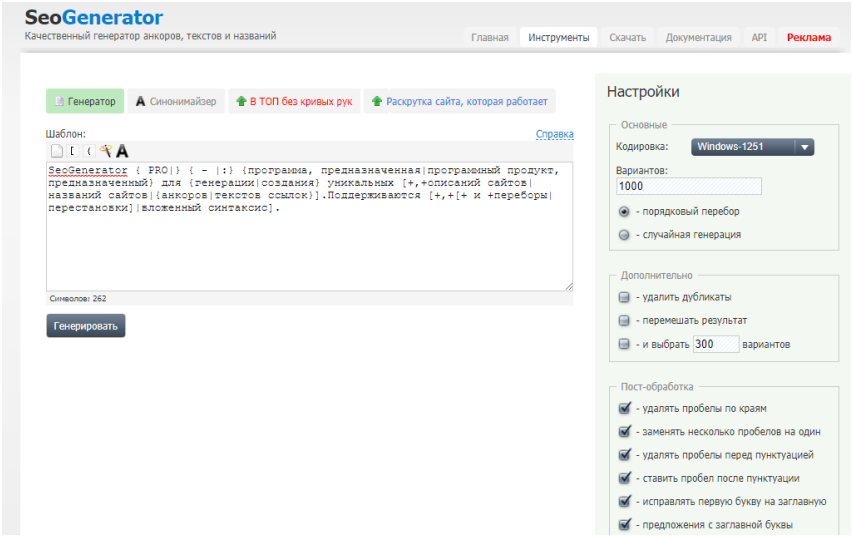

Рисунок 3.1 – Система SeoGenerator

Наприклад, система DRAFTER (рис. 3.2) призначена для створення проектів інструкцій французькою та англійською мовами.

Система містить три компонента – базу знань і два процесорні компоненти: інтерфейс з автором і засоби генерації. База знань, у свою чергу, складається з трьох частин:

1) верхня модель – це онтологія, що описує об’єкти, процеси і властивості, їхні відмінності;

2) модель предметної області – багаторівнева структура, на вищому рівні якої кодуються поняття і відношення, загальні для всіх інструкцій, а на самому нижньому – специфічні для певної предметної області об’єкти та операції;

3) імена конкретних процедур і описів, що згадуються в інструкції, автор з’єднує їх з одиницями двох верхніх рівнів [4, с. 193–197].

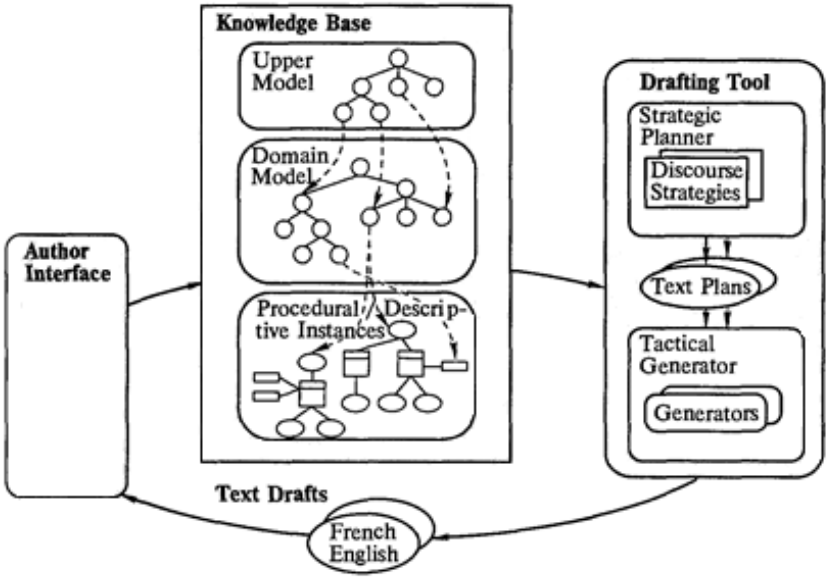

Рисунок 3.2 – Схема процесу генерації тексту

Актуальним завданням залишається багатомовна генерація, тобто автоматична побудова відразу на декількох мовах спеціальних документів – патентних формул, інструкцій з експлуатації технічних виробів або програмних систем, виходячи з їхньої формальної специфікації.

3.4. Створення діалогових систем

Мовне спілкування – це складна діяльність, яка включає в себе безліч процесів від генерації висловлювань до розуміння природно-мовних текстів. При створенні діалогових систем повинні бути прийняті до уваги такі аспекти:

1) порядок обміну репліками (turn-taking);

2) загальний контекст співрозмовників (grounding);

3) структура діалогу (conversational structure);

4) хто бере на себе провідну роль у бесіді (initiative).

Перший аспект – *порядок обміну репліками* – задає той момент, коли наступний учасник повинен вступити в розмову і взяти ініціативу на себе. Зазвичай люди інтуїтивно розуміють, коли прийшов час говорити, але автоматичним системам зрозуміти це буває важко. Логічно припустити, що люди починають говорити, коли настає тиша, але дослідження показують, що паузи нескінченно малі, а люди можуть здогадатися за контекстом, коли їм слід вступати в розмову. Тому розуміння того, що було сказано, відіграє у даному процесі важливу роль. Існують певні правила, за якими можна визначити момент, коли приходить час вступати іншому співрозмовнику. Незважаючи на складність проектування цього аспекту в діалогових системах, він є важливим фактором природності діалогу.

Наявність контексту діалогу – це друга характеристика, яку слід враховувати при розробці діалогових систем. При спілкуванні людям необхідно мати загальний контекст, інформацію, яку співрозмовники використовують для інтерпретації висловлювань один одного.

Важливою характеристикою контексту можна назвати наявність у бесіди теми, яка протягом бесіди може залишатися незмінною або змінюватися. Загальна тема забезпечує розуміння співрозмовниками один одного і в деякій мірі визначає те, як будується діалог.

Ще одним важливим аспектом діалогу є його *специфічна структура*. Вона включає в себе не тільки зміст реплік, а також ролі співрозмовників і що кожен з них повинен повідомити. Дослідження показали, що у людей є внутрішнє розуміння структури бесіди, і вони можуть легко розрізняти природні діалоги і ті, які були автоматично згенеровані. У багатьох випадках структура бесіди зумовлюється ситуацією і диктує зміст. Наприклад, бесіду прийнято починати з вітання, а не з прощання. Таким чином, якщо мета співрозмовників – успішно спілкуватися, вони повинні слідувати всім цим правилам і брати до уваги структуру бесіди.

Поняття *ініціативи в діалозі* є також вкрай важливою характеристикою. Воно показує, хто веде бесіду, має провідне положення. Зазвичай присутня змішана ініціатива (mixed-initiative), де провідна роль перехо-

дить від одного співрозмовника до іншого. Хоча бувають ситуації, коли ініціатива належить тільки одній стороні (single-initiative), наприклад, допит поліцією або суддею.

Всі ці аспекти діалогу дуже важливі при розробці діалогових систем, завдяки їм діалог здається природним і наближеним до людського спілкування.

Сучасні повномасштабні діалогові системи мають більш складну архітектуру і складаються з декількох модулів. Їхня архітектура сильно варіюється, але прийнято вважати, що в ній зазвичай присутні п'ять головних елементів:

1) розпізнавання мови (speech recognition);

2) розуміння мови (natural language understanding);

3) діалоговий менеджмент (dialogue management);

4) генерація природної мови (natural language generation);

5) синтез мови (speech synthesis).

Розпізнавання і синтез мови зазвичай реалізуються як окремі незалежні модулі.

Модуль розуміння природної мови. Завданням модуля розуміння природної мови (natural language understanding) є отримання семантичного представлення вхідного тексту, яке потім зможе бути використано для реалізації будь-яких завдань у рамках діалогових систем.

Для успішної роботи цього модуля необхідна велика кількість інформації, що включає знання про навколишній світ, контекст діалогу та інформацію про дискурс.

Діалоговий менеджер – це центральна складова діалогових систем, яка координує діяльність інших компонентів. Її основним завданням є збереження уявлення про поточний стан діалогу. Головні завдання діалогового менеджера:

– оновлення контексту діалогу на підставі спілкування, що інтерпретується;

– надання контекстно-залежних інтерпретацій сигналів;

– робота з обробкою завдання і області знання (наприклад, бази даних, планування, реалізація та інші системи), координування діалогової і недіалогової поведінки, прийняття рішення про те, яким буде зміст наступного висловлювання і коли слід його реалізувати.

Таким чином, діалоговий менеджер контролює всю архітектуру і структуру бесіди, а також є сполучною ланкою між модулями розуміння і генерації мови, так як бере на себе функцію донесення інформації від одного модуля до іншого.

Модуль генерації природної мови. Генерація природної мови відповідальна за автоматичне створення природно-мовних висловлювань на основі отриманого подання нелінгвістичної інформації.

Приклади діалогових систем:

– Voice Mate (LG);
– S-Voice (Samsung);
– Google Now (Microsoft);
– Cortana (Microsoft);
– Siri (Apple);
– Skyvi (Android) та ін. [3, с. 235–244].

Діалогові системи та чат-боти стають все більш привабливими для комерційного застосування, оскільки вони можуть значно скоротити витрати на обслуговування клієнтів, надавати ефективний доступ до інформації.

Поліпшення діалогових систем, чат-ботів, в тому числі за рахунок застосування машинного навчання і нейронних мереж, створення метрики якості діалогових систем – це перспективні і багатообіцяючі напрямки наукових досліджень і практичних розробок.

3.5. Розпізнавання та синтез мовлення

Додатки, в яких використовуються синтез і розпізнавання мовлення, надзвичайно різноманітні. Це численні комп’ютерні програми, які застосовують голосове введення і виведення інформації, телекомунікаційні

інформаційно-довідкові системи і колл-центри, вхідні дзвінки, які автоматично обробляються і можуть надати на запит користувача необхідні відомості, служби інформування на транспорті, в громадських місцях, за телефоном, різноманітні діалогові системи. Сюди ж можна віднести фонетичні тренажери, які використовуються для поставлення вимови в навчанні іноземної мови, при виправленні дефектів мовлення або при відновленні мовних навичок, а також пристрої, що допомагають людям з обмеженими фізичними здібностями спілкуватися не тільки з комп'ютером, але і з зовнішнім світом. Окремою проблемою є ідентифікація людини за голосом або підтвердження особистості мовця за звукозаписом його мовлення.

Розпізнавання мовлення необхідно для переведення голосового повідомлення в текст. ***Синтез мовлення*** вирішує зворотну задачу. Його мета полягає в породженні мовоподібного звукового сигналу за текстовим записом.

Синтез мовлення. Основне завдання синтезу мовлення полягає в тому, щоб передавати людині голосове повідомлення виключно за допомогою пристрою або механізму, без участі оператора.

За своїм призначенням системи синтезу мовлення належать до однієї з двох груп:

1) предметно-орієнтовані додатки;

2) системи синтезу «від тексту до мовлення».

Останні часто називаються TTS-синтезаторами (від англ. Text-To-Speech).

Предметно-орієнтована група додатків використовує обмежений словник і наперед задані шаблони породжуваних повідомлень, а завданням TTS-систем є озвучування будь-якого тексту за його письмовим представленням.

Методи синтезу. Синтез мовлення, при якому склеюються попередньо підготовлені цифрові звукозаписи, називається *компілятивним* синтезом, або синтезом на основі конкатенації. Для нього необхідно заздалегідь

створити корпус звукозаписів, що містить всі елементи, необхідні для синтезованих фраз.

Альтернативний підхід до породження синтезованого мовлення називається *параметричним* синтезом, або «синтезом за правилами». Це породження нового звукового сигналу «з нічого», приблизно так, як породжується електронним синтезатором звук, що імітує різні музичні інструменти. Тут виділяються такі основні технології: формантний, статистичний, артикуляційний синтез і деякі інші.

Пристрій TTS-синтезатора мовлення. У найзагальнішому вигляді основні компоненти системи синтезу можна подати у вигляді блок-схеми (рис. 3.3):

1) вхідний текст подається на модуль лінгвістичної обробки тексту, який працює з символьною інформацією. В результаті орфографічний запис перетвориться в транскрипційний згідно з існуючими в мові фонетичними правилами;

2) отримана фонетична транскрипція передається на блок синтезу, який і породжує звуковий сигнал за підготовленим транскрипційним записом.

Рисунок 3.3 – Схема синтезу мовлення

Розпізнавання мовлення. Завдання розпізнавання мовлення складається в переведенні в текстовий вигляд (розшифровку) довільного фрагмента усного мовлення або звукозапису.

У загальному вигляді процес розпізнавання виглядає таким чином (рис. 3.4):

1) на вхід системи подається звуковий сигнал або звукозапис;

2) оцифрований звук перетворюється в спектральне представлення, що відображає його об’єктивні акустичні властивості;

3) виділяються важливі для системи акустичні ознаки (вектори);

4) ці вектори порівнюються з шаблонами ознак, які були закладені в пам’яті системи в процесі її навчання;

5) за результатами цього порівняння видається найбільш ймовірний текстовий результат.

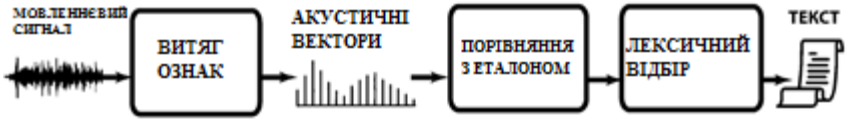

Рисунок 3.4 – Схема розпізнавання мовлення

В цілому, на варіативність мовлення впливають такі чинники:

– анатомічні особливості мовця;

– набуті навички спілкування;

– дефекти мовлення;

– фізіологічний або емоційний стан мовця;

– технічні особливості каналу прийому і передачі інформації;

– навколишнє оточення (шум, одночасне мовлення декількох людей), а також деякі інші.

Лінгвістичний і статистичний підходи до розпізнавання мови. Методи, що використовуються в системах розпізнавання мови, більшою мірою залежать від типу, способу і завдання системи розпізнавання. В цілому всі підходи можна умовно розділити на дві групи – лінгвістичні та статистичні.

Лінгвістичні модулі обов’язково присутні в будь-якій системі розпізнавання. Для їхньої побудови використовуються моделі представлення та обробки мовлення, які виконуються на різних рівнях: акустичному, фонетичному, фонологічному, лексичному, синтаксичному.

Для більш розвинутих систем розпізнавання і розуміння мовлення залучаються вищі рівні аналізу – семантичний і прагматичний.

Однак при розширенні обсягу словника переведення спектральної інформації в послідовність звуків (слів) стає досить складним. Звузити коло пошуку правильного варіанта ідентифікації дозволяють статистичні методи, що враховують ймовірність появи того чи іншого елемента (звуку, слова, словосполучення) в потоці мовлення.

Системи розпізнавання мовлення, що діють на основі статистики, сьогодні вважаються найбільш ефективними. Найчастіше в них використовуються приховані Марковські моделі.

Сучасні системи розпізнавання мовлення розрізняються за обсягом словника, за їхньою прив'язкою до конкретного диктора, а також за типом об'єктів розпізнавання, стилем мовлення, що аналізується, і деякими іншими факторами.

За обсягом словника (тобто за кількістю слів, які вони здатні розрізняти) виділяються такі типи: системи з малим словником розпізнають одиниці або десятки слів, із середнім словником – сотні слів, з великим – тисячі і десятки тисяч слів. В останні роки стали використовуватися поняття «надвеликого словника» для сотень тисяч і навіть мільйонів слів, і «необмеженого словника», завданням якого є моделювання не тільки всіх існуючих, але і потенційно можливих слів для певної мови.

Найпростіші системи орієнтовані на розпізнавання окремих ключових слів або коротких фраз (команди, паролі). Далі за складністю йдуть системи, що розпізнають послідовності, які складаються з обмеженого набору слів (наприклад, числові послідовності, що формують ідентифікаційні коди чи номери телефонів). Нарешті, найбільш складні завдання пов'язані з розпізнаванням зв'язного мовлення або читання зв'язного тексту, диктування і спонтанного мовлення [3, с. 96–120].

Паралельно з поліпшенням якості технологій синтезу і розпізнавання мовлення йде робота з вдосконалення здатності обчислювальної машини як співрозмовника правильно розуміти людську мову – визначати

основну тему мовного повідомлення, його ключові елементи, а в деяких додатках – навіть емоційний або фізичний стан мовця.

3.6. Класифікація та кластеризація

Класифікація (категоризація) – віднесення документа до однієї з декількох заздалегідь визначених категорій на основі змісту документа.

Кластеризація – виявлення груп (їхню кількість заздалегідь не визначено) семантично схожих документів серед заданої фіксованої множини документів.

Можна виділити три основні групи методів класифікації і кластеризації текстів.

По-перше, іноді класифікація або кластеризація здійснюється повністю *вручну*. Наприклад, у бібліотеці, коли бібліотекар присвоює книгам тематичні рубрики.

Інший підхід полягає в написанні правил, за якими можна віднести текст до тієї чи іншої групи. Спеціаліст, який знайомий з предметною областю і володіє навиком написання *регулярних виразів*, може скласти ряд правил, які потім автоматично застосовуються до документів для їхньої класифікації.

Третій підхід ґрунтується на *машинному навчанні*. Для вирішення завдання класифікації документів застосовуються методи машинного навчання з учителем (навчання за прецедентами). Згідно з ними набір правил або критерій прийняття рішення текстового класифікатора обчислюється автоматично з навчальних даних, тобто здійснюється навчання класифікатора.

Наприклад, у програмі електронної пошти може існувати можливість позначати листи як спам, тим самим формуючи навчальну множину для класифікатора – фільтра небажаної пошти.

Таким чином, класифікація текстів, заснована на машинному навчанні, є прикладом навчання з учителем, де в ролі вчителя виступає людина, що задає фіксований набір класів і розмічає навчальну множину.

Завдання кластеризації документів вирішують за допомогою методів машинного навчання без учителя. В цьому випадку не потрібна попередня розмітка частини документів, і кількість класів теж ніяк не фіксується.

Таким чином, розв'язання ***задачі класифікації*** складається з 4 послідовних етапів:

1. Передобробка та індексація документів.

2. Зменшення розмірності простору ознак.

3. Побудова та навчання класифікатора за допомогою методів машинного навчання (метод найменших квадратів, метод k-NN (метод найближчих сусідів); метод опорних векторів; наївний Байєсівський класифікатор та ін.).

4. Оцінка якості класифікації.

Передобробка та індексація документів. Попередня обробка тексту включає в себе токенізацію, видалення функціональних слів (семантично нейтральних слів, таких як сполучники, прийменники, артиклі тощо). Далі здійснюється морфологічний аналіз (виробляється розмітка за частинами мови і стематизація). Це дозволяє значно скоротити розмірність простору. У результаті як ознаки документа виступають всі значущі слова, що зустрічаються в документі. Залежно від програми може індексуватися як весь текст, так і окремі його частини: заголовок, анотація, ключові слова або фрази, перші кілька рядків розділу та ін.

Індексація документів – побудова деякої числової моделі тексту. Вона переводить текст у зручне для подальшої обробки представлення.

Наприклад, модель bag-of-words дозволяє подати документ у вигляді багатовимірного вектора слів та їхніх ваг у документі.

Зменшення розмірності простору ознак. Функція зважування дозволяє оцінити вагу ознаки, визначити, наскільки термін є значущим, або, що те ж саме, виявити ключові слова у цьому документі. Ваги термінів зазвичай нормалізують, щоб вони варіювалися від 0 до 1.

Побудова і навчання класифікатора за допомогою методів машинного навчання. Найбільш поширені методи класифікації:

– метод найменших квадратів (метод регресійного аналізу);

– метод найближчих сусідів (метричний метод класифікації);

– метод опорних векторів (лінійний метод класифікації);

– метод Байєса (імовірнісний метод класифікації);

– метод дерев рішень (логічний метод класифікації).

Оцінка якості класифікації. Основним критерієм при оцінці якості класифікації є комбінація точності і повноти.

Точність (precision) класифікації в межах класу – це частка знайдених класифікатором документів, що дійсно належать даному класу, щодо всіх документів, які система віднесла до цього класу.

Повнота (recall) класифікації – це частка знайдених класифікатором документів, що дійсно належать класу, щодо всіх документів цього класу в тестовій вибірці.

Оцінка якості роботи класифікатора проводиться на тестовій вибірці. Разом з тим роботу системи оцінює експерт (табл. 3.1).

Таблиця 3.1 – Оцінка якості класифікації

Клас c_i		Експертна оцінка	
		Позитивна	Негативна
Оцінка системи	Позитивна	TP	FP
	Негативна	FN	TN

У таблиці застосовані такі умовні позначення:

- TP – істинно позитивне рішення;
- TN – істинно негативне рішення;
- FP – помилково позитивне рішення;
- FN – помилково негативне рішення.

Точність визначається таким чином:

$$p = \frac{TP}{TP + FP}.$$

Повнота обчислюється за формулою:

$$r = \frac{TP}{TP + FN}.$$

Приклади систем автоматичної класифікації текстів:

– Webcat;

– ABBYY Smart Classifier;

– uClassify та ін.

Наприклад, у системі Webcat (рис. 3.5) класифікація заснована на статистичному методі з реалізацією машинного навчання на основі принципу нейронної мережі.

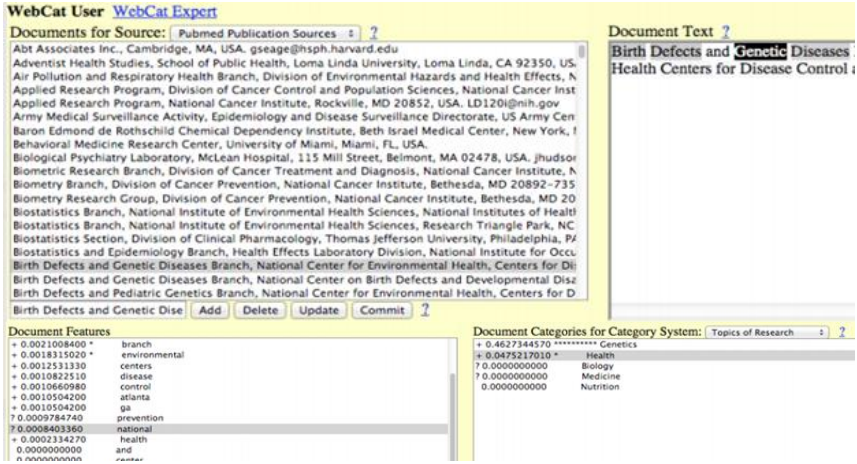

Рисунок 3.5 – Система Webcat

Задача кластеризації полягає в тому, щоб розбити вибірку на непересічні підмножини (кластери) так, щоб кожен кластер складався з об’єктів, близьких за деякою метрикою, а об’єкти різних кластерів істотно відрізнялися.

Для розв'язання задачі кластеризації застосовують *методи навчання без учителя*, коли система навчається виконувати поставлене завдання без стороннього втручання, і не потрібно надавати їй зразки передбачуваного розбиття множини документів на групи. Основні з цих методів:

– ієрархічні методи (метод середнього зв'язку та ін.);

– центроїдні методи (метод *k*-середніх та ін.);

– імовірнісні методи (ЕМ-алгоритм та ін.);

– методи на основі систем штучного інтелекту (метод нечіткої кластеризації С-середніх, нейронні мережі та ін.).

Автоматичні методи класифікації і кластеризації дозволяють обмежити області пошуку в пошукових системах, а отже, підвищити швидкість і точність пошуку, коли вибирається свідомо більш релевантна підмножина документів і виключається свідомо менш релевантна.

Крім того, ці методи застосовуються для складання інтернет-каталогів, підбору контекстної реклами, фільтрації спаму, визначення кодування і мови тексту та ін. [5, с. 113–132].

3.7. Витяг інформації

Основним завданням ***Information Extraction*** є автоматичне екстрагування значущих для людини даних (наприклад, про будь-яку подію), як правило, з великого масиву текстів, і перетворення їх в структуровану форму, що полегшує їх подальшу обробку та аналіз.

Даними, що витягають із текстів, зазвичай виступають:

– значущий об'єкт: ім'я персоналії, назва компанії тощо для новинних повідомлень, термін предметної області спеціального тексту, посилання на літературу для науково-технічних документів і т.д.;

– атрибути об'єкта, що додатково характеризують його, наприклад, для компанії – це юридична адреса, телефон, ім'я керівника і т.п.;

– відношення між об'єктами: наприклад, відношення «бути власником» пов'язує компанію та персону-власника, «бути частиною» з'єднує факультет і університет;

– подія/факт, що зв'язує декілька об'єктів, наприклад, подія «пройшла зустріч» включає учасників зустрічі, а також місце і час її проведення.

Відповідно до видів інформації, що витягується, спільне завдання добування інформації з текстів включає такі основні підзадачі:

– розпізнавання і витяг іменованих сутностей (named entities);

– виділення атрибутів (attributes) об'єктів і семантичних відношень (relations) між ними: дати народження персони, відношення «працювати в» і т.д.;

– витяг фактів і подій (events), що охоплюють кілька їхніх параметрів (атрибутів), наприклад, подія «аварія корабля» з атрибутами дата, час, місце і т.п.

Для вирішення завдання видобування інформації з текстів використовуються два основні підходи:

1) заснований на правилах (rule-based), або інженерний;

2) заснований на машинному навчанні (machine learning).

Відзначимо, що з'являється все більше гібридних методів, що враховують переваги обох підходів.

Інженерний підхід спирається на той факт, що інформація, яка видобувається, використовується в рамках певних мовних конструкцій. Наприклад, назва міста пишеться з великої літери і нерідко передує словами *місто*, або *м*. Подібна лінгвістична інформація зазвичай вручну описується у вигляді формальних шаблонів розпізнавальних конструкцій і правил їхньої обробки. Потім правила застосовуються IE-системою до аналізованого тексту: в ньому шукаються описані шаблонами фрагменти, з яких витягується шукана інформація. Наприклад, за правилом:

ЯКЩО за словом місто, місто-курорт, місто-музей, місто-герой, або м. слідує слово з великої літери,

ТО витягти це слово як назву міста.

У рамках підходу, заснованого ***на машинному навчанні***, застосовуються методи навчання з учителем (supervised), методи навчання без учителя (unsupervised), методи часткового навчання з учителем (bootstrapping).

Найчастіше застосовується навчання з учителем, яке передбачає побудову математичної і програмної моделі, яка вміє відрізняти шукані дані від всіх інших. Побудова такого машинного класифікатора (тобто навчання моделі) відбувається на спеціально розміченому текстовому корпусі (навчальній вибірці), в якому значущим об'єктам, їхнім атрибутам, відношенням, фактам приписані відповідні мітки. Мітки кодують ознаки для розпізнавання цих даних. Для наведеного вище прикладу для вилучення назви міста як ознаки можуть виступати: регістр (верхній) першої літери слова, конкретні слова, які стоять перед ним (місто, місто-курорт, місто-музей, місто-герой, або м.), а також ознаки таких слів (для виявлення багатослівних назв).

Загалом, навчання моделі полягає у виявленні на основі окремих даних, які увійшли до навчальної вибірки, загальних закономірностей.

Приклади систем витягу інформації:

– GATE;
– OpenNLP;
– Eureka Engine;
– Stanford CoreNLP та ін.

Наприклад, Stanford CoreNLP включає:

1. Модуль Stanford Named Entity Recognizer для витягу іменованих сутностей.

2. Модулі Stanford Relation Extractor і Stanford OpenIE для виявлення відношень.

3. Модуль Stanford Pattern-based Information Extraction and Diagnostics для ітеративної побудови набору шаблонів для витягу інформації.

4. Модуль Stanford Relation Extractor орієнтований на пошук відношень між двома сутностями. Навчена для англійської мови модель дозволяє витягувати 4 типи семантичних відношень:

– «жити в», що пов'язує ім'я персони і географічний об'єкт;

– «знаходитися в», що пов'язує два географічні об'єкти;

– «розташовуватися в», що пов'язує назву організації та географічний об'єкт;

– «працювати в», що пов'язує ім'я персони і назву організації [2, с. 83–121].

Іншим прикладом системи автоматичного виявлення інформації з текстів є Eureka Engine (рис. 3.6). Один з модулів автоматичного визначення іменованих сутностей реалізований у вигляді сервісу, який дозволяє класифікувати іменовані об'єкти в тексті на п'ять класів (фізичні особи, юридичні особи, географічні об'єкти, назви продуктів і брендів та іменовані події).

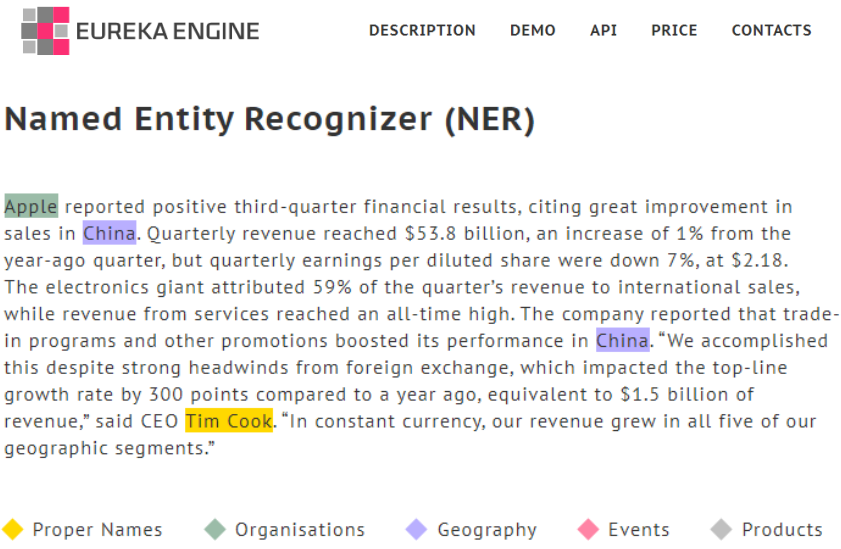

Рисунок 3.6 – Система Eureka Engine

Таким чином, сучасними тенденціями розвитку цього напрямку є:

– розширення використання різних факторів і ресурсів, зокрема, великих зовнішніх ресурсів знань (Вікіпедія, DBPedia, WordNet, графи знань та ін.);

– урахування при виявленні нелокальних залежностей текстових одиниць;

– проведення більш глибокого синтаксичного аналізу та використання синтаксичних ознак при машинному навчанні [2, с. 121].

3.8. Аналіз тональності текстів

Автоматичний аналіз тональності текстів – виявлення думки автора тексту з приводу предмета, що обговорюється в тексті, є однією з технологій, що активно розвиваються в сфері автоматичної обробки текстів [2].

При аналізі тональності необхідно виявити кілька складових:

– джерело думки, або суб'єкт тональності – хто є автором повідомлення;

– об'єкт тональності – про що йде мова в тексті;

– аспект тональності – характеристика об'єкта;

– тип думки (оціночний компонент), або тональна оцінка. Тут мається на увазі ставлення автора до описуваного предмета, конкретне повідомлення про аспекти (властивості) об'єкта.

Підходи аналізу тональності:

1. *Побудова правил.* У правилах для аналізу тональності використовують заздалегідь розроблені шаблони, що описують певну предметну область. За цими шаблонами з тексту витягуються *n*-компонентні ланцюжки (*n*-грами), їхня тональність визначається як на основі правил, так і на основі словників.

2. *Підхід з використанням машинного навчання*. При навчанні з учителем алгоритм класифікації тренується на основі навчальної вибірки. Цю вибірку потрібно зібрати і розмітити [3, с. 247–255].

Сучасні словникові ресурси для аналізу тональності:

– MPQA;
– SentiWordNet;
– WordNet-Affect;
– SenticNet;
– AFINN та ін.

Наприклад, AFINN створено для аналізу постів у соціальних мережах, які включають сленгові слова. Він містить близько 2 400 слів, позначених числовою вагою полярності, що змінюється від –5 (дуже негативний) до +5 (дуже позитивний):

abandon –2, abduction –2, abhor –3, abusive –3, accept 1 ...

Таким чином, словники оціночної лексики, створені для однієї мови, значною мірою різняться між собою за покриттям, а також можуть відрізнятися за приписаними оцінками тональності для конкретних слів [2, с. 135–138].

Прикладом системи аналізу тональності коротких неструктурованих текстів є SentiStrength. Результат роботи системи видається у вигляді двох оцінок – оцінка позитивної складової тексту (за шкалою від +1 до +5) і оцінка негативної складової (за шкалою від –1 до –5).

Системи автоматичного аналізу тональностей активно розвиваються. Актуальність пов'язана з розвитком соціальних мереж, онлайнових рекомендаційних сервісів, що містять велику кількість думок людей з різних питань, зокрема, про різні товари, послуги.

Завдання

Навести приклади сучасних застосунків комп'ютерної лінгвістики. Описати підходи і методи, які в них використовуються.

Література

1. Соснина Е. П. Введение в прикладную лингвистику / Е. П. Соснина. – Ульяновск : УлГТУ, 2012. – 110 с.

2. Автоматическая обработка текстов на естественном языке и анализ данных / Е. И. Большакова, К. В. Воронцов, Н. Э. Ефремова, Э. С. Клышинский, Н. В. Лукашевич, А. С. Сапин. – Москва : Изд-во НИУ ВШЭ, 2017. – 269 с.

3. Николаев И. С. Прикладная и компьютерная лингвистика / И. С. Николаев, О. В. Митренина, Т. М. Ландо. – Москва: Ленанд, 2016 . – 316 с.

4. Леонтьева Н. Н. Автоматическое понимание текстов: системы, модели, ресурсы : учеб. пособ. / Н. Н. Леонтьева. – Москва : ИЦ «Академия», 2006. – 304 с

5. Батура Т. В. Математическая лингвистика и автоматическая обработка текстов / Т. В. Батура. – Новосибирск : РИЦ НГУ, 2016. – 166 с.

6. Барсегян А. А. Анализ данных и процессов / А. А. Барсегян, М. С. Куприянов, И. И. Холод, М. Д. Тесс, С. И. Елизаров. – Санкт-Петербург: БХВ-Петербург, 2009. – 512 с.

7. Маннинг К. Д. Введение в информационный поиск: пер. с англ. / К. Д. Маннинг, Р. Прабхакар, Ш. Хайнрих. – Москва : ООО «И.Д. Вильямс», 2011. – 528 с.

8. Обробка тексту [Електронний ресурс]. – Режим доступу : https://nlpub.ru/Обработка_текста.

9. Computational Linguistics and Intelligent Text Processing: Lecture Notes in Computer Science / edited by A. Gelbukh. – Springer, 2018. – 664 p.

10. Deng L. Deep Learning in Natural Language Processing / L. Deng, Y. Liu – Springer, 2018. – 329 p.

РОЗДІЛ 4
ПРИКЛАДНІ АСПЕКТИ КОРПУСНОЇ ЛІНГВІСТИКИ

4.1. Створення корпусу текстів.

4.2. Розмітка корпусів.

4.3. Застосування корпусних менеджерів.

4.4. Використання корпусів.

4.1. Створення корпусу текстів

Корпусна лінгвістика – розділ комп'ютерної лінгвістики, що займається розробкою загальних принципів побудови і використання лінгвістичних корпусів (корпусів текстів) із застосуванням комп'ютерних технологій. Під лінгвістичним, або мовним, ***корпусом текстів*** розуміється великий, поданий у машинопрочитуваному форматі, уніфікований, структурований, розмічений, філологічно компетентний масив мовних даних, призначений для вирішення конкретних лінгвістичних завдань.

До поняття «корпус текстів» входить також система управління текстовими та лінгвістичними даними, яку останнім часом найчастіше називають ***корпусним менеджером***.

Пошук у корпусі дозволяє за будь-яким словом побудувати ***конкорданс*** – список усіх вживань певного слова в контексті з посиланнями на джерело.

Процес створення корпусу можна подати у вигляді таких кроків або етапів:

1. Забезпечення надходження текстів відповідно до переліку джерел.

2. Перетворення в машинопрочитувану форму.

3. Аналіз і попередня обробка текстів.

4. Конвертація і графематичний аналіз. Деякі тексти проходять також через один або кілька етапів попередньої машинної обробки, в ході яких здійснюється перекодування (якщо потрібно), а також видалення або перетворення нетекстових елементів (рисунки, таблиці), видалення з тексту переносів, забезпечення однакового написання тире і т.д. Графема-

тичний аналіз передбачає проведення таких операцій: поділ вхідного тексту на елементи (слова, роздільники і т.д.), видалення нетекстових елементів, виділення та оформлення нестандартних (нелексичних) елементів, обробка спеціальних текстових елементів (імен, написаних ініціалами, іноземних лексем, записаних латиницею, назв рисунків, приміток, закреслень, титульних аркушів, списків літератури і т.д.).

5. Розмітка тексту. Розмітка тексту полягає в приписуванні текстів та їхніх компонентів додаткової інформації (метаданих). Метадані можна поділити на 3 типи:

– екстралінгвістичні, що належать до всього тексту;
– дані про структуру тексту;
– лінгвістичні метадані, що описують елементи тексту.

Метаопис текстів корпусу включає як змістовні елементи даних (бібліографічні дані, ознаки, що характеризують жанрові та стильові особливості тексту, відомості про автора), так і формальні (ім'я файлу, параметри кодування, версія мови розмітки, виконавці етапів робіт). Ці дані зазвичай вводяться вручну. Структурна розмітка документа (виділення абзаців, речень, слів) і лінгвістична розмітка зазвичай здійснюються автоматично.

6. Коригування результатів автоматичної розмітки: виправлення помилок і зняття неоднозначності.

7. Конвертація розмічених текстів у структуру спеціалізованої лінгвістичної інформаційно-пошукової системи (corpus manager), що забезпечує швидкий багатоаспектний пошук і статистичну обробку.

8. Забезпечення доступу до корпусу.

9. Створення документаційного забезпечення, в якому описуються різні аспекти створення і використання корпусу, зокрема, наводяться відомості про розмітку, що дозволяють шукати за метаданими, мова запитів корпус-менеджера і т.д.

4.2. Розмітка корпусів

Фактично, корпус в його сучасному розумінні – це завжди комп'ютерна база даних, і в процесі його створення використовуються спеціальні процедури і програми. Наприклад, токенізація, тобто розбиття потоку символів у природній мові на окремі значущі одиниці (токени, словоформи), є необхідною умовою для подальшої обробки природної мови.

Інша специфічна задача морфологічного аналізу – це лематизації, тобто процес утворення первісної форми слова, виходячи з інших його словоформ.

Процес, дещо відмінний від лематизації, називається стемінг, він складається зі знаходження стема (основи) слова. Різниця полягає в тому, що стемер обробляє окреме слово без знання контексту, і, таким чином, не може диференціювати слова, які мають різні значення в силу віднесеності до різних частин мови.

Парсинг – це процес зіставлення лінійної послідовності лексем (слів, токенів) мови з її формальною граматикою. Результатом зазвичай є дерево залежностей (синтаксичне дерево).

Серед спеціальних програм для обробки природної мови особливе місце займають програми автоматичної розмітки.

Розмітка полягає в приписуванні текстам та їхнім компонентам спеціальних тегів: лінгвістичних, що описують лексичні, граматичні та інші характеристики елементів тексту, і зовнішніх, екстралінгвістичних (відомості про автора і відомості про текст: автор, назва, рік і місце видання, жанр, тематика).

Серед ***лінгвістичних типів розмітки*** виділяються: морфологічна, синтаксична, семантична, анафорична, просодична, дискурсна та ін.

Морфологічна розмітка (POS-tagging). Насправді морфологічні мітки включають не тільки ознаку частини мови, а й ознаки граматичних категорій, властивих конкретній частині мови.

Синтаксична розмітка є результатом парсинга, що виконується на основі даних морфологічного аналізу. Цей вид розмітки описує синтаксичні зв'язки між лексичними одиницями і різні синтаксичні конструкції.

Семантична розмітка. Семантичні теги найчастіше позначають семантичні категорії, до яких належить певне слово або словосполучення, і вужчі підкатегорії, що специфікують його значення. Семантична розмітка корпусів передбачає специфікацію значення слів, вирішення омонімії і синонімії, категоризацію слів (розряди), виділення тематичних класів, ознак каузативності, оціночних і деривaційних характеристик і т.д.

Існують також інші типи розмітки, зокрема:

– *анафорична розмітка* фіксує референтні зв'язки, наприклад, займенників;

– *просодична розмітка.* У просодичних корпусах застосовуються теги, що позначають наголос та інтонацію. У корпусах усного мовлення просодична розмітка часто супроводжується так званою дискурсною розміткою, яка служить для позначення пауз, повторів, обмовок і т.д.

Екстралінгвістична розмітка, або метадані, включає в себе «зовнішню», «інтелектуальну» розмітку (бібліографічні характеристики, типологічні характеристики, тематичні характеристики, соціологічні характеристики), «формальну» структурну розмітку (текст, розділ, абзац, речення), а також техніко-технологічну розмітку (кодування, дату обробки, виконавців, джерело електронної версії). Набір метаданих багато в чому визначає можливості, що надаються корпусами дослідникам.

4.3. Застосування корпусних менеджерів

Корпусний менеджер – спеціалізована пошукова система, що включає програмні засоби для пошуку даних у корпусі, отримання статистичної інформації і надання результатів користувачеві в зручній формі.

Корпусний менеджер повинен:

– будувати конкордансні списки;

– шукати не тільки окремі слова, а й словосполучення;

– здійснювати пошук за шаблонами (складні запити);

– сортувати списки за кількома критеріями, що обираються користувачем;

– давати можливість відображати знайдені словоформи в необмеженому контексті;

– давати статистичну інформацію з окремих елементів корпусу;

– відображати леми, морфологічні характеристики словоформ і метадані (бібліографічні, типологічні), що залежать від ступеня розмічування корпусу;

– зберігати і роздруковувати результати;

– працювати як з окремими файлами, так і з корпусами, необмеженими за розміром;

– швидко обробляти запити і видавати результати;

– підтримувати різні формати текстових даних (txt, doc, rtf, html, xml та ін.);

– бути легким (інтуїтивно зрозумілим) у використанні як для досвідченого користувача, так і для початківців.

Найбільш відомі такі універсальні корпусні менеджери:

– SARA;

– XAIRA (BNC);

– Manatee / Bonito;

– CQP;

– DDC та ін.

Для обробки корпусних даних можуть розроблятися менеджери на основі систем управління базами даних або пошукових систем [1, с. 7–56].

Більшість сучасних корпусних менеджерів включають статистичні методи, багатомовну підтримку та ін. Приклади таких інструментів:

– WordSmith Tools;

– MonoConc Pro;

– ParaConc;

– AntConc та ін.

Для роботи з корпусами невеликих розмірів часто використовується AntConc (рис. 4.1). Це вільне програмне забезпечення, зі зручним інтерфейсом, що має безліч функцій з обробки текстів.

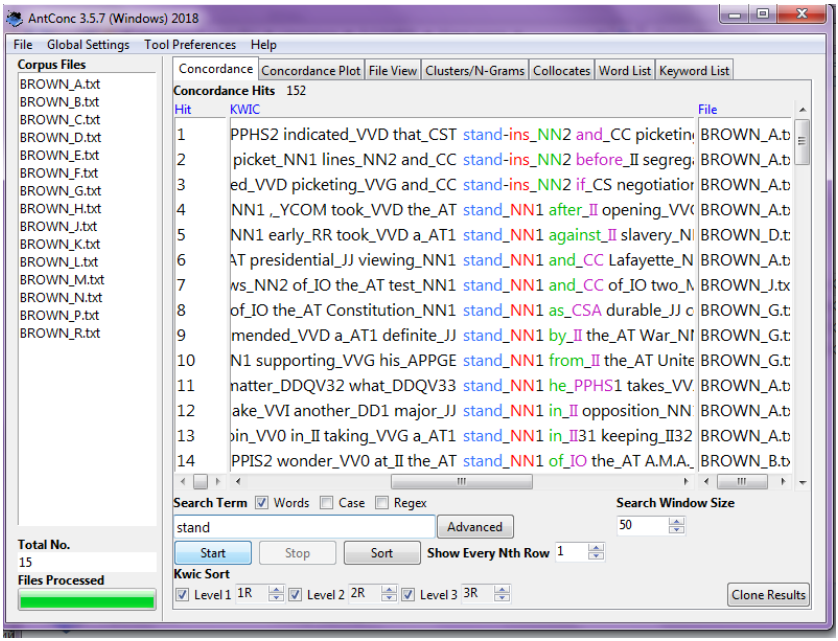

Рисунок 4.1 – Сисема AntConc

За допомогою AntConc можна проводити такі операції:

– перегляд файлу з текстом;

– побудова конкордансу для заданого слова в межах контекстного вікна;

– побудова графіків до конкордансу;

– виділення *n*-грам зі заданим словом у межах контекстного вікна;

– виділення колокатів заданого слова на основі мір асоціації;

– побудова частотного списку словоформ і / або лем із зазначенням рангу і абсолютної частоти;

– виділення ключових слів [2, с. 151].

4.4. Використання корпусів

Багато лінгвістів використовують корпус як «банк прикладів», тобто намагаються знайти емпіричну підтримку для своїх гіпотез, принципів і правил, над якими вони працюють.

Емпірична підтримка являє собою якісний метод використання корпусу, але корпуси також забезпечують інформацію з частотності для слів, фраз і конструкцій, яка може бути використана для кількісних досліджень.

На додаток до лінгвістичного контексту корпус надає екстралінгвістичну інформацію, або *метаінформацію*, за такими факторами, як вік або стать мовця, жанр тексту, тимчасова або просторова інформація про походження тексту і т.д. Вона дозволяє порівнювати різні типи текстів чи різні групи мовців.

Лексикографічні дослідження необхідні, в першу чергу, для складання словників.

Граматичні дослідження, засновані на корпусах: вивчення граматики пов'язано з розумінням структури мови, включаючи морфологію, синтаксис та інші властивості слів, такі як їхні граматичні класи (дієслова, іменники та ін.).

Приклади існуючих корпусів текстів:

– *Корпус сучасної американської англійської мови* включає тексти, що являють собою усне мовлення, художню прозу, популярні журнали, газети та наукову літературу обсягом 425 млн слів.

– *Корпус DeReKo (das Deutsche Referenz Korpus)* складається з белетристики, наукових і публіцистичних текстів німецької мови і містить понад 4,1 млрд слововживань.

– *Британський національний корпус* містить більше 100 млн слів розмовної та письмової англійської мови. Підкорпус письмової мови становить 90 % всього корпусу і включає в себе газети, періодичні наукові видання і журнали для різних вікових груп, популярну наукову фантастику, опубліковані і неопубліковані листи, шкільні та університетські твори та ін. Підкорпус усного мовлення включає в себе множину контекстів: від

мови формальних ділових або урядових зустрічей до радіошоу і телефонних розмов.

– *Чеський національний корпус* – синхронічний морфологічно розмічений корпус, який представляє сучасну чеську мову. Масив текстів, накопичених у корпусі, ділиться на синхронічну і діахронічну частини. Синхронічна частина складається з таких частин: письмові тексти (понад 100 млн слововживань), розмовні (750 тис. слововживань) і діалектна мова. Обсяг діахронічної частини – 1 750 тис. слововживань.

– *Національний корпус української мови* – вибірка текстів сучасної української мови, обсягом близько 50 млн слововживань. Корпус містить прозові, поетичні та драматургічні тексти відповідних хронологічних меж і метатекстову розмітку.

– *Національний корпус російської мови* – інформаційно-довідкова система, заснована на російських текстах, обсягом близько 343 млн слововживань. Корпус містить тексти наукової, офіційно-ділової, публіцистичної, церковно-богословської, художньої, розмовно-побутової спрямованості.

– *Корпус усних оповідань російською мовою* містить 595 елементарних дискурсивних одиниць, які зазвичай збігаються з простих речень, і 327 ілюстративних жестів, які являють собою носії інформації, виступаючи як знакові кінетичні одиниці виразу і передачі інформації.

– *Російськомовний емоційний корпус*, розмічений з урахуванням даних про міміку, рухи рук, брів і т.д., дозволяє вивчити стратегії емоційної взаємодії та конфлікту, безперервну комунікативну поведінку, хезітації та мовні збої та ін. [1, с. 75–115].

Таким чином, великі національні корпуси дозволяють дослідникам здійснювати автоматичний пошук і систематизацію емпіричного матеріалу, швидко обробляти великі масиви мовних даних за допомогою спеціальних комп'ютерних програм.

Корпусна лінгвістика являє собою одну з областей, що стрімко розвивається. У свою чергу, прогрес у сфері комп'ютерних технологій тягне за собою прогрес у створенні і вдосконаленні програм автоматичної обробки тексту.

Завдання

Вказати функції сучасних корпусних менеджерів.

Література

1. Захаров В. П. Корпусная лингвистика / В. П. Захаров, С. Ю. Богданова. – Иркутск : ИГЛУ, 2011. – 169 с.

2. Николаев И. С. Прикладная и компьютерная лингвистика / И. С. Николаев, О. В. Митренина, Т. М. Ландо. – Москва: Ленанд, 2016. – 316 с.

3. Щипицина Л. Ю. Информационные технологии в лингвистике / Л. Ю. Щипицина. – Москва : ФЛИНТА : Наука, 2013. – 128 с.

4. Desagulier G. Corpus Linguistics and Statistics with R: Introduction to Quantitative Methods in Linguistics / G. Desagulier. – Springer; 2017. – 353 p.

РОЗДІЛ 5
ЗАСТОСУВАННЯ СУЧАСНИХ ТЕХНОЛОГІЙ У ЛЕКСИКОГРАФІЇ

5.1. Класифікація словників.

5.2. Макро- та мікроструктура словника.

5.3. Комп'ютеризація словників.

5.1. Класифікація словників

Лексикографію визначають як теорію і практику складання словників.

Словник – це певним чином організоване зібрання слів, зазвичай з приписаними їм коментарями, в яких міститься інформація про особливості структури і / або функціонування лексичних одиниць.

У сучасній лексикографії виділяють два напрямки діяльності: теоретичний та практичний. Теоретичний напрямок займається теорією та історією лексикографії, а практична лексикографія – це безпосередня діяльність зі створення словників.

Практична лексикографія виконує кілька функцій:

1) навчальну функцію – словники полегшують вивчення мови як рідної, так і нерідної;

2) функцію нормалізації мови – словники (у першу чергу тлумачні, орфографічні, орфоепічні) описують і нормалізують мову;

3) комунікативну функцію – словники (двомовні словники, розмовники) забезпечують міжмовне спілкування;

4) дослідницьку функцію – дані для досліджень черпаються з етимологічних, частотних, зворотних та інших спеціальних словників.

Теоретична лексикографія охоплює ряд проблем:

1) розробка загальної типології словників;

2) розробка макроструктури словника (відбір лексики, принцип розташування слів і словникових статей та ін.);

3) розробка мікроструктури словника, тобто окремої словникової статті (включення фонетичного та граматичного коментарів, позначок, спеціальних знаків та ін.).

У сучасній лексикографії існує декілька класифікацій словників. В. В. Дубічинський [1] запропонував узагальнюючу типологію словників:

1. Першою підставою класифікації стала ***кількість мов***, що описуються. Виділяють одномовні:

– пояснювальні (енциклопедичні, тлумачні, термінологічні, ономастичні, навчальні тощо);

– фіксуючі, що представляють списки слів без пояснень (орфографічні, частотні, зворотні, ідеографічні, орфоепічні).

Одномовним словникам протиставляються двомовні (багатомовні) або перекладні словники.

2. За ***охопленням лексики***:

– словники, в які лексика включена без обмежень (енциклопедичні, тлумачні, орфографічні, орфоепічні, тезауруси);

– словники, які описують тільки певні лексичні пласти (словники неологізмів, діалектів та ін.).

3. За ***обсягом*** словника виділяють три групи:

– великі чи повні;

– стислі;

– лексичні мінімуми.

4. За ***оформленням і деталізацією інформації***:

– комп'ютерні;

– друковані (багатотомні, однотомні, кишенькові, ілюстративні).

5. За ***функціональною спрямованістю***:

– функціонально-галузеві (термінологічні, тематичні, тезауруси);

– функціонально-мовні (словники сполучуваності слів);

– функціонально-образні (фразеологічні).

6. Наступний параметр – ***порядок подачі*** лексичного матеріалу:

– алфавітні (тлумачні, орфографічні, перекладні та ін.);

– ономасіологічні (тезауруси, ідеографічні словники та ін.);

– алфавітні зворотні (словотвірні, граматичні).

7. За тематикою (ономастичні, країнознавчі та ін.).

8. Змішані / комплексні словники (наприклад, тлумачно-перекладні).

9. Навчальні словники.

Ця класифікація не включає в себе кілька типів словників, наприклад, словники, одиниці опису яких менше слова, тобто словники морфем, асоціативні словники, а також словники скорочень.

5.2. Макро- та мікроструктура словника

Макроструктура словника. Кожен словник складається з ряду компонентів. Набір цих компонентів утворює макроструктуру словника.

Прийнято виділяти ліву і праву частини словника. ***Ліва частина*** – це словник (вхід словникових статей). ***Права частина*** – опис одиниць словника: тлумачення або переклад, а також граматична, стилістична, етимологічна і словотвірна інформація.

На відбір лексики впливає існуюча літературна норма. Більшість лексикографів сходиться на трьох критеріях норми:

– представленість мовного факту у найбільш авторитетних авторів;

– поширеність явища;

– вимога відповідності явища основним законам мови.

На етапі відбору матеріалу лексикограф повинен визначити, які джерела він буде використовувати. Зазвичай джерелами стають письмові тексти, усні висловлювання носіїв мови, раніше видані довідники і словники. При створенні словників активно використовуються корпуси текстів.

Крім словника, макроструктуру словника становить ***передмова***, де пояснюються загальні принципи побудови словника, даються вказівки щодо його використання. Там же пояснюється структура словникової статті.

Ще однією важливою складовою частиною лінгвістичного словника є ***список умовних скорочень*** і ***алфавіт***.

Мікроструктура словника. Елементарною одиницею будь-якого словника є ***словникова стаття*** – кожен окремо взятий об'єкт опису словника і словникові характеристики, які відносяться до цього об'єкта.

Словникова стаття складається з декількох зон опису, кожна з яких містить особливий тип словникової інформації. Кількість зон і характер інформації залежить від типу словника.

Перша зона словникової статті – ***лексичний вхід*** (вокабула або лема). Часто в лемі вказується наголос.

Слідом за лексичним входом йде ***зона граматичної інформації*** та ***зона стилістичних позначок***. Надається граматична інформація про приналежність слова до частини мови, вказуються особливі граматичні форми. Стилістичні позначки вказують на сферу вживання слова.

Словникова стаття, як правило, має наступну структуру:

1. Лексичний вхід.
2. Стилістичні позначки.
3. Граматична інформація.
4. Тлумачення.
5. Приклади вживання.
6. Зона ідіоматики (стійкі словосполучення з певною одиницею, фразеологізми).

У двомовних словниках словникова стаття, як правило, має таку структуру:

1. Лема.
2. Зона фонетичної інформації.
3. Зона граматичної інформації.
4. Зона еквівалентів.
5. Відсильна зона.

5.3. Комп'ютеризація словників

Комп'ютерна лексикографія – напрямок лексикографії, який займається розробкою електронних словників.

Для створення словників використовуються спеціальні програми – бази даних, корпуси текстів та ін. Множину лексикографічних програм можна розділити на:

- програми підтримки лексикографічних праць;
- автоматичні словники різних типів.

Основні завдання при створенні електронного словника:

1) визначення форми заголовної одиниці;

2) визначення обсягу інформації, яка включатиметься до опису лексичної одиниці;

3) вибір способу організації словникової статті.

Структура словникової статті в різних електронних словниках може виглядати по-різному, обсяг статті може коливатися від 1 до 99 зон, в яких фіксується різнотипова лінгвістична і екстралінгвістична інформація:

- заголовне слово;
- один або кілька еквівалентів (у перекладному словнику);
- інформація про тематичну приналежність слова;
- граматична інформація;
- дефініція;
- контексти;
- інформація про лексичну сполучуваність;
- семантичні ієрархічні зв'язки заголовного слова;
- інформація про стилістичні характеристики слова;
- джерело та ін.

Наприклад, словникова стаття електронної версії *The New Oxford Dictionary of English* виглядає таким чином:

<se> standard entry, *or*
<ee> encyclopedic entry, *embedding*
<hw> headword
<pr> pronunciation
<s1> sense level 1 (part of speech)
<ps> part of speech
<s2 num-n> sense level 2, with number attribute, *embedding*
<df> definition
<ms> meaning extention
<ex> example of usage (taken from *The British National Corpus* or *The Oxford English Dictionary* citation files)
<et> etimology
<drv> derivation form, *embedding*
<ps> part of speech [2, с. 4–55].

Окремі електронні словники мають також додаткові можливості. Наприклад, електронний багатомовний словник ABBYY Lingvo надає функцію навчання ABBYY Lingvo Tutor (рис. 5.1).

ABBYY Lingvo Tutor дозволяє запам'ятовувати слова, відібрані за конкретною темою, складати нові словники та словникові картки, зберігати результати навчання в файлі.

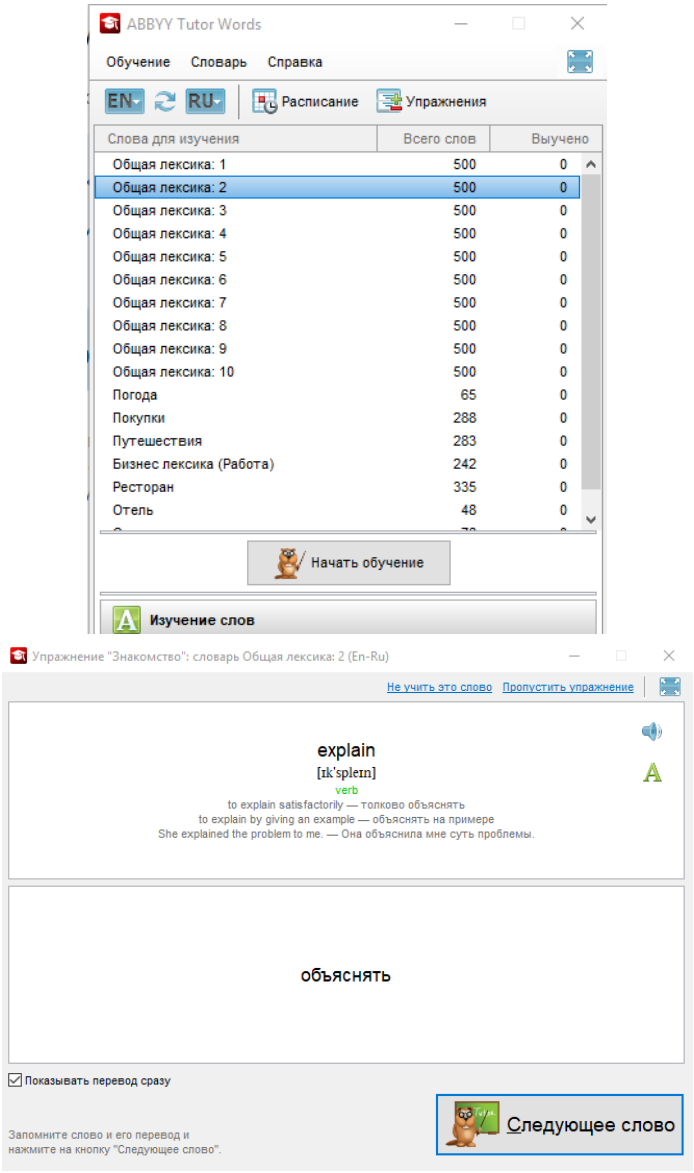

Рисунок 5.1 – Функція ABBYY Lingvo Tutor

Завдання

Описати структуру сучасного електронного словника, вказавши:

- структуру словникової статті;
- базу словників (типи, тематики, обсяг);
- додаткові функції, при наявності.

Література

1. Дубічинський В. В. Українська лексикографія: історія, сучасність і комп'ютерні технології / В. В. Дубічинський. – Харків : НТУ «ХПІ», 2004. – 203 с.

2. Шилихина К. М. Теоретическая и практическая лексикография / К. М. Шилихина. – Воронеж : ВГУ, 2006. – 59 с.

3. Щипицина Л. Ю. Информационные технологии в лингвистике / Л. Ю. Щипицина. – Москва : ФЛИНТА : Наука, 2013. – 128 с.

4. Electronic lexicography in the 21st century: linking lexical data in the digital age // Proceedings of the eLex 2015 conference. – edited by I. Kosem, M. Jakubíček, J. Kallas, S. Krek. – Brighton, United Kingdom, 2015. – 517 p.

РОЗДІЛ 6
МІЖДИСЦИПЛІНАРНІ ДОСЛІДЖЕННЯ В ЛІНГВІСТИЦІ

6.1. Когнітивна лінгвістика.
6.2. Психолінгвістика.
6.3. Нейролінгвістика.
6.4. Комунікативна лінгвістика.
6.5. Соціолінгвістика.

6.1. Когнітивна лінгвістика

Когнітивна лінгвістика – галузь мовознавства, яка вивчає мову як засіб отримання, зберігання, обробки, переробки та використання знань, спрямована на дослідження способів концептуалізації і категоризації світу дійсності і внутрішнього рефлексивного досвіду. *Об'єктом* когнітивної лінгвістики є мова як показник когнітивних структур і процесів свідомості, а *предметом* – співвідношення когнітивних механізмів свідомості з природною мовою та її мовної реалізації.

Головним джерелом когнітивної лінгвістики вважається *комп'ютерна наука* та її лінгвістична галузь, спрямована на розробку автоматизованих методів зберігання, обробки, переробки та використання лінгвістичних знань та інформації, представленої знаками природної мови.

Другим, не менш важливим, джерелом когнітивної лінгвістики є *когнітивна психологія* – галузь експериментальної психології, спрямована на вивчення пізнавальних процесів людини: сприйняття, пам'яті, уваги, мислення, планування діяльності, навчання і т.д.

Третім джерелом когнітивної лінгвістики став *генеративізм*. Когнітивна лінгвістика несе відбиток деяких ідей теорії породжувальної граматики, зокрема, уявлення про мову як породжувальний механізм, ментальну репрезентацію граматики окремого носія мови, моделювання цих ментальних процесів.

Головними *завданнями* когнітивної лінгвістики є:

1) аналіз природи мовної компетенції людини, її онтогенезу. ***Мовна компетенція*** – це шлях до можливості якісно і продуктивно говорити іншою мовою в різних сферах діяльності відповідно до норм мови, що вивчається;

2) визначення специфіки категоризації і концептуалізації досвіду в колективній свідомості носіїв мови;

3) опис організації внутрішнього лексикону, вербальної пам'яті людини відповідно до структур репрезентації знань і механізмів пам'яті взагалі;

4) пояснення когнітивної діяльності людини в процесах породження, сприйняття і розуміння мови, комунікативної діяльності;

5) дослідження пізнавальних процесів і ролі природних мов в їх здійсненні;

6) встановлення співвідношення мовних структур з когнітивними і т.п. [1, с. 365–377].

Напрямки в когнітивній лінгвістиці:

– *культурологічний* – дослідження концептів як елементів культури, спираючись на дані різних наук. Такі дослідження зазвичай міждисциплінарні, не пов'язані виключно з лінгвістикою, хоча можуть виконуватися і лінгвістами (що і дозволяє розглядати даний підхід в рамках когнітивної лінгвістики); мова в цьому випадку виступає лише як одне з джерел знань про концепти (наприклад, для опису концепту використовуються дані про етимологію слова, що називає цей концепт);

– *лінгвокультурологічний* – дослідження названих мовними одиницями концептів як елементів національної лінгвокультури в їхньому зв'язку з національними цінностями і національними особливостями цієї культури: напрямок «від мови до культури»;

– *логічний* – аналіз концептів логічними методами поза прямої залежності від їхньої мовної форми;

– *семантико-когнітивний* – дослідження лексичної і граматичної семантики мови як засобу доступу до змісту концептів, як засобу їх моделювання від семантики мови до концептосфери;

– *філософсько-семіотичний* – досліджуються когнітивні основи знаковості.

Матеріалом лінгвокогнітивного аналізу є мова, а цілі такого дослідження в різних конкретних напрямках (школах) когнітивної лінгвістики можуть відрізнятися – від поглибленого вивчення мови за допомогою когнітивного категоріально-термінологічного апарату до конкретного моделювання змісту і структури окремих концептів як одиниць національної свідомості (концептосфери) [2 , с. 9–12].

Концепт – інформаційна структура свідомості, певним чином організована одиниця пам'яті, яка містить сукупність знань про об'єкт пізнання, вербальних і невербальних, набутих шляхом взаємодії п'яти психічних функцій свідомості і несвідомого.

Методи опису концептів. Лінгвокогнітивна методика опису і моделювання концептів будується на аналізі семантики мовних засобів, що об'єктивують концепт. Аналіз семантики (традиційно-лінгвістичний та експериментальний) може дати досліднику як опис значень лексичних одиниць (лексикографічний та психолінгвістичний), так і опис концепту.

Лексикографічне значення формулюється на основі аналізу численних контекстів вживання слова шляхом логічної редукції семантичних ознак, визнаних лексикографом несуттєвими. Якщо слово позначає різні денотати, то виділяються різні значення, які розташовуються в словниковій статті від основного до похідних, більш пізніх за виникненням, пов'язаних з основними відношеннями семантичної похідності.

Методика опису психолінгвістичного значення слова може бути подана таким чином:

1. Проведення асоціативного експерименту з досліджуваним словом як стимулом.

2. Побудова асоціативного поля досліджуваного слова-стимулу.

3. Семантична інтерпретація асоціативних реакцій як мовних репрезентацій сем.

4. Семна атрибуція отриманих сем (розподіл сем за окремими значеннями за денотативним принципом).

5. Семний опис змісту і структури виділених значень як зв'язкової сукупності сем.

6. Моделювання семантеми як впорядкованої множини виділених сем за принципом убування яскравості семи в семантемі.

Приклад опису значення слова *«жертва»*:

1. *Асоціативний експеримент*. Інформанти навпроти запропонованого слова-стимулу повинні записати перше слово, яке прийде їм в голову. У разі якщо таке слово не знаходиться, опитувані ставлять прочерк. Отримані відповіді обробляються і будується асоціативне поле стимулу «жертва».

2. *Побудова асоціативного поля*. Асоціативне поле досліджуваного слова:

Жертва 500 – вбивство 33, злочин 30, насильство 26, труп 23, потерпілий 19, хижак 14, біль, смерть 12, кров 11, вівця 10, терор 8, жалість, любов, слабкість, теракт 7, обставини 6, життя, маніяк, потерпілий, секта, страх 5, ...

3. *Семантична інтерпретація асоціативних реакцій*. Семантична інтерпретація отриманих реакцій полягає в осмисленні отриманих асоціацій. Далі на цьому ж етапі здійснюється формулювання сем, тобто таке формулювання семантичних компонентів, яке дозволяє приписати цю сему значенням досліджуваного слова. Отримані формулювання сем повинні узгоджуватися з досліджуваним словом, наприклад, *жертва*: є результатом злочину 114, є наслідком обставин 52, є результатом слабкості 11, є результатом переконання 11, викликає жалість 13 і т.д.

4. *Семна атрибуція отриманих сем.* На цьому етапі семи поділяються на групи за кількістю окремих значень, які вони представляють в смисловій структурі слова:

1) жива істота 29 (вівця 10, життя 5, тварина 4, птиця, ягня 3, свиня 2, заєць 1); кого вбивають 19 (кров 11, позбавлення життя 2, тварина, яку вбивають 1 і т.д.); ...

2) добровільна жертва 7: вклад, милостиня 3, дар 1;

3) добровільна відмова від кого/чого-небудь на чиюсь користь 11:

мета 6: заради чогось 2; …

оцінка 3: дурість, марно, необхідність 1;

4) самопожертва 5: самовіддача 4, дієта 1;

5) причина 183: злочин 114 (вбивство 33, злочин 30, насильство 26, терор 8, теракт 7, напад 4, садизм, тероризм 2, пограбування, злочинність 1); обставини 52 (війна, любов 7, обман 4, репресії 3, автокатастрофа, ...

5. *Семний опис значень*. Кожна сема отримує дефініцію, яку дослідник словесно формулює:

Жертва 340:

1) той, хто постраждав у результаті злочину 114, обставин 52, слабкості 11, своїх переконань 3; суб'єкт, що зазнає вплив, є трупом 23, потерпілий 19, ...

2) жива істота 29, принесена 7 в дар божеству 14 для очищення гріхів 3;

3) відмова 2 заради чогось 3 або когось 1 з визначеною метою 1; є дурістю 2, необхідністю 1;

4) пожертвування 7;

5) самопожертва 5.

6. *Моделювання семантеми слова*. За кількістю випробовуваних, які актуалізували різні значення досліджуваного слова, можна надати ієрархію значень у смисловій структурі досліджуваного слова, розраховуючи індекс яскравості значення як відношення числа випробовуваних, які актуалізували компоненти цього значення в експерименті, до загальної кількості випробовуваних.

Основним значенням слова *«жертва»* буде «той, хто постраждав в результаті чогось» (індекс яскравості 78,2 %). Саме воно і утворює ядро

семантеми досліджуваного слова. Ближню периферію становить значення «істота, яку ритуально убивають» (індекс яскравості 16,6 %). Дальня периферія подана значенням «відмова від чогось або когось» (індекс яскравості 2,5 %). Крайню периферію утворюють значення «пожертвування» (індекс яскравості 1,6 %) і «самопожертва» (індекс яскравості 1,1 %).

Моделювання концепту включає три процедури, які взаємодоповнюються, але виконуються окремо:

1) опис макроструктури концепту (віднесення виявлених когнітивних ознак до образного, інформаційного компонентів та інтерпретаційного поля і встановлення їхнього співвідношення в структурі концепту);

2) опис категоріальної структури концепту (виявлення ієрархії когнітивних класифікаційних ознак, що концептуалізують відповідний предмет або явище, і опис концепту як їхньої ієрархії за актуальністю для концептуалізації денотата концепту);

3) опис польової організації концепту (виявлення та опис когнітивних класифікаційних ознак, що становлять ядро, ближню, дальню та крайню периферію концепту і подання змісту концепту у вигляді польової структури).

1. *Опис макроструктури концепту*. Процедура передбачає розподіл виділених когнітивних ознак за структурними макрокомпонентами концепту – образною, інформаційною складовою та інтерпретаційним полем. Це дозволяє наочно подати, які типи інформації переважають у концепті і яке їхнє співвідношення один з одним.

Наприклад, концепт *«англійська мова»* за результатами спрямованого асоціативного експерименту має такий макроструктурний склад (відсоток вказується від загального числа отриманих реакцій – 690):

Образний компонент – 20 %:

– яскравий, зоряно-смугастий, Біг Бен (перцептивні образи) – 7 %);

– розумний, тупий, ввічливий, культурний (когнітивні метафори) – 13 %.

Інформаційний зміст – 35 %:

– всесвітній, Америка, Британія.

Інтерпретаційне поле – 45 %:

– оціночна зона: прекрасний, хороший, чужий, жахливий, красивий, цікавий, грубий, оригінальний;

– утилітарна зона: загальнодоступний, складний;

– регулятивна зона: необхідний;

– соціально-культурна зона: сучасний, бітл, Біг Бен.

2. *Опис категоріальної структури концепту*. Наприклад, асоціативний експеримент зі словом *«борг»* і подальша когнітивна інтерпретація результатів експерименту дозволяють виділити такі когнітивні ознаки, що утворюють зміст концепту *борг*:

– обов'язок 89, відповідальність 37, моральна тяжкість 10, гроші 75, необхідність повернення 21, перед Батьківщиною 35, перед рідними, перед роботою 3, перед другом 4, перед людством, законом 1, через потреби 1, через карти 2, веде до розплати 3, дуелі, помсти 1, виконання приносить почесті, покликання 1, буває священний, цивільний, академічний 1, носій друг, лицар, собака 1.

Дані когнітивні ознаки в структурі концепту об'єднуються, інтегруються в такий спосіб:

– *складові* 214: обов'язок 89, відповідальність 37, почуття 2; гроші 75, необхідність повернення 21;

– *наслідки* 17: тяжкість 10, веде до розплати 3, дуелі, помсти, виконання приносить почесті, покликання 1;

– *сфера прояву* 50: перед Батьківщиною 35; перед рідними, перед роботою 3, перед другом 4, перед людством, перед законом 1; …

– *причина появи* 3: нужда, картковий, карти 1;

– *видові різновиди* 3: священний, цивільний, академічний;

– *носій* 3: один, лицар, собака 1.

3. *Опис польовий організації концепту*. Підсумком моделювання польової організації концепту в рамках лінгвоконцептологічного дослі-

дження є словесне або графічне подання змісту концепту у вигляді польової структури.

Як приклад наведемо графічне подання (рис. 6.1) польової організації концепту *«жінка»* [2, с. 116–152].

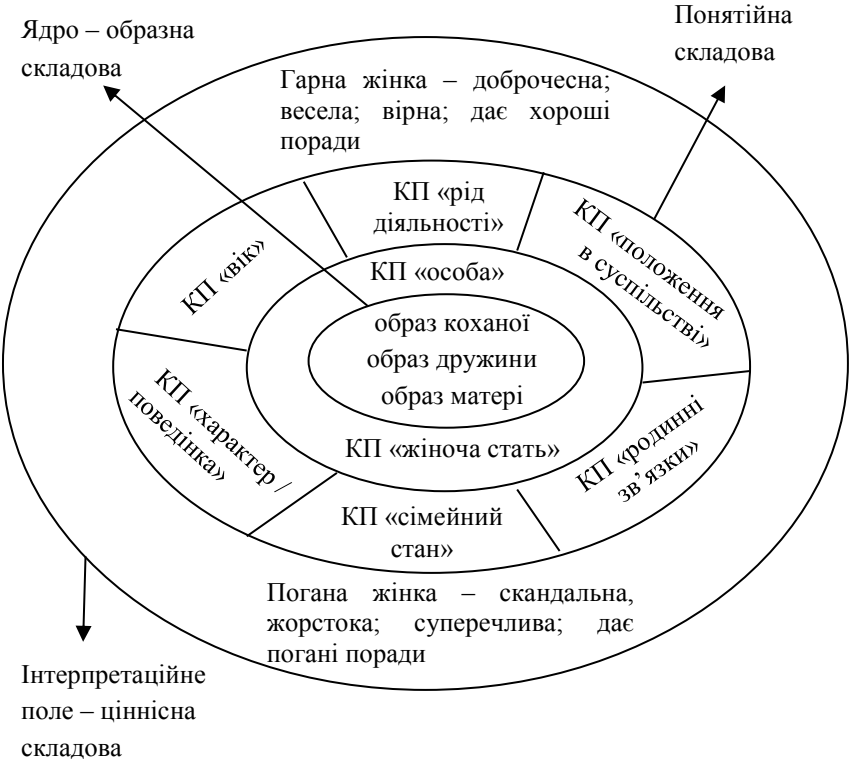

Рисунок 6.1 – Приклад опису польової організації концепту «жінка»

Таким чином, в завдання когнітивної науки входить і опис/вивчення систем подання знань і процесів обробки та переробки інформації, і – одночасно – дослідження загальних принципів організації когнітивних

здібностей людини в єдиний ментальний механізм, і встановлення їхнього взаємозв'язку і взаємодії.

Кінцевим завданням когнітивної лінгвістики, як і когнітивної науки в цілому, є отримання даних про діяльність розуму [3, с. 8–13].

6.2. Психолінгвістика

Психолінгвістика – наука, що вивчає психологічні та лінгвістичні аспекти мовної діяльності людини, соціальні та психологічні аспекти використання мови в процесах мовної комунікації та індивідуальної мовленнєво-розумової діяльності.

Предметом дослідження психолінгвістики є мовна діяльність як специфічно людський вид діяльності, її психологічний зміст, структура, види (способи), в яких вона здійснюється, форми, в яких вона реалізується, функції, які нею виконуються.

Об'єктом дослідження психолінгвістики виступають: людина як суб'єкт мовної діяльності і носій мови, процес спілкування, комунікації.

Методи дослідження психолінгвістики можна розділити на три великі групи:

- загальна методологія;
- спеціальна (тобто конкретно-наукова) методологія;
- спеціальні (конкретно-наукові) дослідні методи.

Загальною методологією є філософія, що розуміється як світогляд.

Спеціальну методологію становлять закони науки, її теорія, гіпотези, наукові концепції, аксіоми і поняття, методологічні принципи і т.д.

Можна виділити 4 групи *дослідних методів,* які використовуються в психолінгвістиці: організаційні, емпіричні, обробні, інтерпретаційні.

За допомогою ***організаційних методів*** виконується психолінгвістичне дослідження закономірностей формування і здійснення мовної діяльності.

До ***емпіричних методів*** належать:

– об'єктивне спостереження. Так, дослідження обмовок дозволяє виявити багато специфічних властивостей мовних процесів, а також випадків мовної поведінки, які цікавлять дослідників;

– бесіда, анкетування, опитування, тести і ряд інших.

Методи обробки – різноманітні статистичні методи, методи опису отриманих даних дослідження.

Таким чином, психолінгвістика нерозривно пов'язана:

– з філософією, яка сприяє загальному напрямку дослідження;

– з психологією (загальною, віковою, соціальною, спеціальною психологією і багатьма іншими її областями). Без даних практичної психології психолінгвістика не може бути достатньо обґрунтованою наукою;

– з лінгвістикою (загальним мовознавством, філософією мови, граматикою певної мови, соціолінгвістикою, етнолінгвістикою та іншими розділами лінгвістики);

– з семіотикою – наукою про знаки мови та їхнє значення;

– з логікою;

– з соціологією;

– з медициною, переважно з неврологією, яка чимало сприяла вивченню патологій і норм мови, а також з психіатрією, оториноларингологією та рядом інших медичних наук, з логопатологією, логопедією та іншими науками логопатологічного кола, котрі надають багато цінних даних для розуміння процесів породження і сприйняття мови;

– з деякими технічними науками (зокрема, з тими, які роблять можливим апаратне і комп'ютерне забезпечення досліджень мовленнєвої діяльності та мовних знаків);

– з акустикою і психоакустикою та ін. [4, с. 9–14].

Прикладні аспекти психолінгвістики

Судова психолінгвістика займається аналізом не стільки особливостей мовного повідомлення, скільки мовної (і мовленнєвої) особистості, яка стоїть за текстом.

Можна виділити кілька сфер застосування судової психолінгвістики: мовленнєва взаємодія під час допиту, аналіз показань свідків, визначення істинності (або хибності) висловлювання, ідентифікація особистості за мовленням.

1. *Психолінгвістика допиту*. Одним з розділів судової психології є психологія допиту і показань.

Для судового психолога важливі такі моменти:

1) як зафіксувати показання;

2) які тактики допиту і методи впливу на допитуваного допустимі;

3) як діагностувати завідомо неправдиві показання.

Так, в одному з психолінгвістичних експериментів випробуваним показували документальний фільм, де було зображено зіткнення автомобілів, які рухалися з різною швидкістю – 20, 30 і 40 миль на годину. Після перегляду фільму випробовуваних просили відповісти на питання про швидкість руху автомобілів, і виявилося, що вони були не дуже точні в своїх відповідях, оскільки їхні відповіді були в основному однакові і не залежали від реальної швидкості руху машин.

Потім були взяті дві групи випробовуваних, однорідних у віковому і статевому плані. Після перегляду документального фільму їм задавали подібні питання про швидкість руху машин перед аварією. Але питання ці різнилися між собою одним словом. Відповіді на них виявилися різними залежно від того, яке дієслово використовувалося в питанні: нейтральне або те, що містить експресію (табл. 6.1).

Виявилося, що навіть візуальні враження (що особливо важливо для урахування ступеня точності показань свідків) виявляються під впливом слів, що використовуються в питанні, і тим самим визначають значною мірою характер сприйняття.

Таблиця 6.1 – Приклад психолінгвістичного експерименту

Питання	Відповідь (миль на годину)
How fast were the cars going when they HIT each other?	34,8
How fast were the cars going when they COLLIDED with each other?	
How fast were the cars going when they BUMPED each other?	
How fast were the cars going when they CONTACTED with each other?	
How fast were the cars going when they SMASHED into each other?	40,8

2. *Розпізнавання брехні в мовленні.* Брехня є феноменом мовного спілкування, що складається в навмисному спотворенні дійсного стану речей.

Як показують експерименти і спостереження, мовлення людини, що намагається спотворити реальний стан справ або ввести кого-небудь в оману, відрізняється від мовлення людини, що говорить правду.

При описі вигаданої ситуації немає вступних слів з відтінком припущення і невпевненості, тому що описувана ситуація випробуваному відома. У вигаданих текстах набагато більше елементів, пов'язаних з логічної обробкою інформації (1), на відміну від правдивих текстів, де переважають модальні слова, що позначають можливість (2). Референтні індекси (3) у вигаданих текстах зустрічаються частіше, ніж зазвичай. Невизначені займенники (4) зустрічаються частіше у вигаданих текстах, ніж у правдивих (5). У правдивих текстах більше лексики, що належить до кінестетики і до візуального каналу сприйняття дійсності. При описі вигаданої ситуації більше динаміки, ніж при описі реальної ситуації. Обсяг мовної продукції при брехні менше. Мовлення повільніше, в ньому більше обмовок:

1) це так, тому що;

2) можливо це так, тому що;

3) ну, ти знаєш, це як завжди;

4) який-небудь, де-небудь;

5) якийсь, десь.

Розпізнавання брехливого висловлювання вимагає уваги не лише до голосових характеристик мовлення, а й до жестів і міміки мовця. Велика кількість людей (в тому числі і поліцейські) здатністю до розпізнавання брехні не володіють. Зокрема, глядіння в очі співрозмовника, яке вважається чесним, характерне швидше для того, хто обманює, ніж для того, хто говорить правду. А саме той, хто обманює, зберігає більш тривалий контакт очей зі слухачем, ніж той, хто говорить правду. Положення тіла того, хто обманює, затиснуте і напружене.

Одним із сучасних засобів розпізнавання брехні є поліграф, або детектор брехні. Характерно, що в основу конструювання поліграфа лягли саме дослідження в області психолінгвістики.

Виділяють три типи реакції на подразнення (слабким розрядом електричного струму), а саме:

– нейтральна;

– орієнтовна;

– захисна.

При нейтральній реакції на зовнішній подразник-стимул кровоносні судини пальців і голови знаходяться в звичайному, нормальному стані і не показують жодних змін. При орієнтовній реакції, коли людина активно реагує на якийсь подразник-стимул, але без додаткової мобілізації сил, судини пальців стискаються, в той час як судини голови розширюються. Нарешті, при захисній реакції, коли людина, реагуючи на несприятливий стимул, мобілізує сили організму до опору, судини і пальців, і голови стискаються. Спеціальні датчики чітко фіксують кожне з трьох станів.

Сучасний детектор брехні є апаратурним комплексом, який слугує для об’єктивної реєстрації фізіологічних показників. До них належать: КДР (кардіограма), ЕЕГ (електроенцефалограма), тремор (тремтіння

пальців рук), плетизмограма (графічна реєстрація динаміки кровонаповнення кровоносних судин), пульс, кров'яний тиск, глибина дихання, потовиділення та ін. Ці показники використовуються з метою аналізу емоційних відповідей на стимули, що пред'являються під час бесіди або допиту.

Наприклад, випробуваному задаються нейтральні питання, відповіді на які експериментатору відомі (1), а також емоційно значущі (2).

1. *Ви носите взуття?*

2. *Ви коли-небудь вживали марихуану?*

Деякі труднощі застосування поліграфа пов'язані з тим, що вже сама процедура допиту може викликати сильні емоції (страх, тривогу, депресію), які складно відрізнити від реакції на смисл питань. Тому до числа недоліків діагностичної сили поліграфа відносять також труднощі інтерпретації змісту реакції людини.

3. *Ідентифікація особистості за мовленням.* Судова психолінгвістика може також допомогти у вирішенні проблеми ідентифікації особистості за мовленням.

Одним з досить давно існуючих методів визначення особистості за мовленням є почеркознавчий аналіз, який досліджує особистість за почерком. *Почеркознавча експертиза* – це також спосіб:

– встановлення конкретних виконавців, або виконання рукописів одним або різними особами;

– встановлення умов їхнього виконання (зовнішні і внутрішні чинники);

– встановлення соціально-демографічних параметрів їхніх виконавців (їхня стать і вік).

Відповідно до деяких уявлень почерк як прояв мовленнєвої діяльності людини залежить від профілю або акцентуації особистості. Велику роль відіграють і властивості нервової системи, психофізіологічні особливості особистості. Так, почерк екстраверта може характеризуватися нахилом вправо, а інтроверта – вліво. Сильний натиск може бути характерний для домінантних особистостей, слабкий – для невпевнених у собі. Параноальність (як схильність до утворення надцінних ідей) проявляється не-

рідко в більшій кількості великих літер, більш частому початку з червоних рядків. Епілептоїдність (як характеристика збудливої особистості) виявляється в дрібному почерку, у великій кількості розділових знаків і сильному натиску. Істероїдність (бажання звернути на себе увагу) – це, перш за все завитки, закрутистість (в тому числі і в підписі), «відкриті» букви, хвилястість ліній. Депресивність (пониження настрою) – нахил лінії вниз в кінці рядка, слабкий натиск, маленький міжрядковий інтервал і маленька відстань між буквами, словами і полями. Маніакальність (підвищений настрій) – великі літери. Маніакально-депресивні розлади настрою – це хвилясті рядки. Окремо написані літери можуть свідчити про низьку соціалізацію особистості. Характерно, що особливості почерку зберігаються і коли людина пише іноземною для неї мовою.

У психології розслідувань історично склалися три методи психологічного вивчення особистості, які мають загальний об'єкт дослідження, але різняться за програмними цілями і завданнями, предметом і техніками:

– контактний;
– слідовий;
– дистанційний.

Контактна психодіагностика – це складання психологічного портрета особистості при безпосередній взаємодії з нею. *Слідова психодіагностика* – це складання психологічного портрета особистості по слідах її життєдіяльності (житло, особисті речі, об'єкти впливу та ін.). *Дистанційна психодіагностика* – це опосередковане вивчення особистості за допомогою аудіовізуального спостереження, вивчення фото- і кінодокументів, опитування близьких людей, використання технік непрямого допиту і непрямого обстеження та ін. Важливе місце тут займає аналіз мовлення і текстів, які напевно належать досліджуваній особистості.

Так, при аналізі реального випадку зникнення дружини був лист на шести сторінках, нібито написаний нею. У ньому йшлося про те, щоб її не шукали, оскільки вона «пішла до іншого». Психолінгвістичний аналіз декількох листів чоловіка показав, що в них міститься набагато більше гра-

матичних, синтаксичних, орфографічних і пунктуаційних помилок, ніж у листах, дійсно написаних його дружиною. Такого ж роду помилки містилися і в листі, який підлягав аналізу. Зроблений висновок судового психолінгвіста про підробку «прощального» листа згодом підтвердився визнанням чоловіка у вбивстві.

На сьогоднішній день актуальними є експериментальні дослідження проблеми ідентифікації особистості за усним мовленням і голосом. До основних характеристик голосу належать тембр, висота основного тону, гучність.

За голосом можна дізнатися про такі фізичні стани, як гальмування (втома, важкість, розслабленість, слабкість, млявість, безсилля, депресія, сонливість, стан напівсну, знемога, апатія, байдужість) або збудження (у тому числі хвилювання, нервозність). Можна сказати, чи налаштована людина на виконання якої-небудь дії, тобто чи є в неї впевненість (емоційний підйом, гарний настрій, спокій) або розгубленість (невпевненість, задуманість, байдужість, довірливість, співчутливість, пригніченість, безнадійність, роздум).

За голосом можна дізнатися і про емоційний стан людини, як наприклад, тривогу (переляк, страх, жах), роздратованість (невдоволення, гнів, злість, обурення, лють, ненависть, загрозу, обурення), радість (захоплення, задоволення, захоплення, щастя), сум (смуток, туга, відчай), засмученість (гіркоту, горе, жаль). За голосом можна дізнатися, висловлює оратор насмішку (єхидство, знущання), образу (досаду) або ніжність.

Мова і гендер. Дослідження мови чоловіків й жінок показують, що є певні відмінності між тим, як говорять і пишуть чоловіки, і тим, як говорять і пишуть жінки.

В теперішній час правомірно говорити про певні особливості мовленнєвого стилю чоловіків і жінок. Він проявляється на двох рівнях – мовленнєвої поведінки й мовлення. Так, чоловіки частіше перебивають, більш категоричні, прагнуть керувати тематикою діалогу і т.д. Істотно, що на відміну від поширеної думки, чоловіки говорять більше, ніж жінки. Чоловічі речення, як правило, коротші жіночих. Чоловіки в цілому наба-

гато частіше вживають абстрактні іменники, а жінки – конкретні (в тому числі власні імена). Чоловіки частіше використовують іменники (в основному, конкретні) і прикметники, в той час як жінки вживають більше дієслів. Чоловіки вживають більше відносних прикметників, а жінки – якісних. Чоловіки частіше використовують дієслова доконаного виду в дійсному стані.

Жіноча мова виявляє велику концентрацію емоційно-оцінної лексики, а чоловіча оцінна лексика частіше стилістично нейтральна.

Нейролінгвістичне програмування – набір явно зазначених навичок і технік мовних та немовних взаємодій.

Наприклад, що стосується кінестетики (переважно органів чуттів, а також положення тіла), то її вплив на наше мовлення може бути проілюстровано такою вправою: *Сядьте на стілець так, щоб ноги щільно були притиснуті до підлоги. Дихайте неглибоко і рідко. Нахиліть голову направо, опустіть плечі. Тепер, дивлячись вниз – на ноги, скажіть наступну фразу: «Це - найщасливіший день мого життя».*

Невідповідність фрази, що вимовляється, положенню тіла викликає труднощі її вимовляння, що можна перевірити у такій алогічній вправі: *Встаньте прямо. Дихайте глибоко всією діафрагмою, звільніть ваші плечі і шию, потягніться трохи і трохи посміхайтеся. Роблячи це, говоріть: «Ви уявити собі не можете, яким пригніченим я відчуваю себе прямо зараз».*

Реакцією на спробу вимовити саме цю фразу саме в такому положенні може стати сміх, викликаний неузгодженістю кінестетичного з лінгвістичним.

Таким чином, нейролінгвістичне програмування вивчає суб'єктивний досвід, патерни, або «програми», створені взаємодією мозку (нейро), мови (лінгвістичне) і тіла, які обумовлюють як ефективну, так і неефективну поведінку. Нейролінгвістичні техніки можуть бути використані в різних галузях професійної комунікації, включаючи психотерапію, бізнес, гіпноз, юриспруденцію і освіту. Прийоми нейролінгвістично-

го програмування використовують багато авторів реклами, бізнесмени і політики.

Мова і мовленнєвий вплив. Крім тієї допомоги, яку слова надають як інструменти мислення, вони надають можливість здійснювати вплив. При цьому слова не тільки приносять користь адресату висловлювання, але з ними можуть бути пов'язані зловживання, що містяться в самому акті комунікації. А саме: можливі спотворення сенсу внаслідок зміни загальновідомого значення слів заради маскування реальності, обмеження і навіть «затуманення» мислення людини, зміни її поведінки і дій. Все це можна робити за допомогою мови. Мовне і мовленнєве маніпулювання одержувачами інформації аж до примусу людини діяти всупереч його інтересам – це реальність.

Одним з важливих аспектів мовленнєвого впливу є те, що він здійснюється за допомогою усно наданої інформації. Тому, якщо на певних словах навмисно робиться наголос, якщо мовлення добре структуроване чи сконструйоване зі спеціальною метою, усна інформація може надавати набагато більший вплив, ніж письмова.

Ритмічність мови має особливе значення. Якщо порівняти два описи (1 і 2), то ми побачимо, що другий з них, де є зміна інтонаційного рисунка, виявляється більш легким для засвоєння і більш дієвим у плані зміни стану свідомості слухача (табл. 6.2).

При ритмічному читанні з правильно розставленою інтонацією мовлення змінюється за синусоїдом: з підвищенням інтонації і з пониженням. Вдало підібраний сюжет, на який добре лягає чергування інтонацій, дозволить читачеві ввести слухача в легкий транс, перевантаживши його увагу і психотерапевтичним змістом, і інтонацією. Після цього тексту можуть вже наслідувати власне тексти з відповідною установкою [5, с. 188–216].

Таблиця 6.2 – Приклад мовленнєвого впливу

<table>
<tr><td>1 Мовлення рівномірне, монотонне, без інтонацій</td><td colspan="2">2 У мовленні з’являються паузи, тон голосу по черзі то підвищується, то знижується</td></tr>
<tr><td rowspan="11">Уявіть собі човен, який гойдається на хвилях вгору і вниз, тому що навіть у безвітряну погоду на воді є якісь хвилі, і на цих хвилях човен хитається вгору і вниз, і на воді завжди є сонячні зайчики, які стомлюють ваш зір і серйозно приковують до себе вашу увагу.</td><td>Уявіть собі човен, … який гойдається на хвилях … вгору …</td><td>голос підвищується, йдучи майже в дискант</td></tr>
<tr><td>і вниз ...</td><td>голос знижується до низького баритона</td></tr>
<tr><td>тому що навіть в безвітряну погоду ...</td><td>голос підвищується</td></tr>
<tr><td>на воді є якісь хвилі ...</td><td>голос знижується</td></tr>
<tr><td>і на цих хвилях ...</td><td>голос підвищується</td></tr>
<tr><td>човен хитається …</td><td>голос знижується</td></tr>
<tr><td>вгору ...</td><td>голос підвищується</td></tr>
<tr><td>і вниз ...</td><td>голос знижується</td></tr>
<tr><td>і на воді завжди є сонячні зайчики ...</td><td>голос вище</td></tr>
<tr><td>які стомлюють ваш зір ...</td><td>голос нижче</td></tr>
<tr><td>і серйозно приковують до себе вашу увагу ...</td><td>зовсім тихо</td></tr>
</table>

Таким чином, психолінгвістика має найбільш тісні зв’язки із загальним мовознавством. Крім того, вона постійно взаємодіє з соціолінгвістикою, етнолінгвістикою і прикладною лінгвістикою, особливо з тією її частиною, яка займається питаннями комп’ютерної лінгвістики.

Психолінгвістика – це міждисциплінарна галузь знань про закони формування в онтогенезі і сформованих процесах мовленнєвої діяльності в системі різних видів життєдіяльності людини [4, с. 16].

6.3. Нейролінгвістика

Нейролінгвістика – лінгвістична дисципліна, що виникла на стику неврології (нейрофізіології) і лінгвістики та вивчає психо-фізіологічний механізм мовного відображення дійсності. В останні десятиліття актуальними стали дослідження зворотного зв'язку – вплив мови на протікання фізіологічних процесів.

Об'єктами вивчення є:

1) всі види афазії:

– динамічна, яка має синтагматичну природу, при порушенні передньої мовної зони;

– семантична – труднощі в побудові смислової схеми висловлювання;

– синтаксична – аграматизм, труднощі в граматичній організації висловлювання та еферентна, моторна – труднощі в моторній кінетичній організації мовлення;

2) неафатичні розлади мови (мовні агнозії, апраксія, дізартрії, алексії та опрафії).

Предметом вивчення є кореляції між тими чи іншими видами розладів мови і функціональними центрами кори головного мозку. Інакше кажучи, предметом вивчення є психофізіологічний механізм мовного відображення дійсності (у тому числі розпізнавання реаналізаторів мозку і процеси мовних узагальнень). При цьому враховується функціональна асиметрія півкуль мозку, яка обумовлює переважну локалізацію мовних узагальнень і мислення в мовних поняттях у лівій (домінантній) півкулі, а конкретно-образного мислення – у правій (субдомінантній) півкулі.

На основі синтезу нейролінгвістики, психолінгвістики, прагматики, когнітивної науки і психоаналізу виникла теорія і технологія нейролінгвістичного програмування, *метою* якого є вивчення та застосування способів оптимізації через мовленнєвий вплив функціонування кори головного мозку, що відповідає за свідомість, і центрів, які несуть відповідальність за сферу підсвідомості [6, с. 526–528].

Репрезентативні системи як моделі сприйняття. У нейролінгвістичному програмуванні під ***репрезентативними системами*** маються на увазі індивідуальні моделі сприйняття і прийняття того, що передають нам наші органи чуттів.

Усіх людей фахівці з нейролінгвістичного програмування поділяють на візуалів, аудіалів, кінестетиків і дискретів (табл. 6.3).

«Візуали» зазвичай стоять прямо з прямими плечима або спиною, причому тримають шию теж прямо – відповідно до тіла. Коли вони йдуть, здається, що їх «веде підборіддя», а рухи візуально орієнтованої людини можна охарактеризувати як різкі або рвучкі. У типових «візуалів» ребра видаються меншими, ніж у людей з інших категорій, і вони зазвичай дихають верхньою частиною грудної клітки. Досить часто можна бачити, що візуал говорить швидко, ясно або на більш високих нотах, ніж люди інших типів.

Типові *«кінестетики»* зазвичай бувають більш повними, ніж індивіди з інших категорій (хоча вони далеко не завжди виявляються повними). При спілкуванні вони часто демонструють округлі плечі, а іноді просто нахиляються вперед, коли говорять або слухають. Їхні рухи зазвичай плавні та вільні. «Кінестетики» мають більш визначні ребра у порівнянні з людьми інших типів і зазвичай дихають нижньою частиною легенів. Тональність голосу кінестетично орієнтованих суб'єктів узагальнено може бути подана як м'яка і повітряна, а мовлення відзначається повільним темпом, низьким тоном і гучністю.

Фізична конституція *«аудіалів»* швидше тонка ніж повна. Звичайна їх комунікативна поза – руки складені на грудях, а голова нахилена вниз і вбік ніби слухаючи. Коли «аудіали» кажуть, то приділяють значно більше уваги, ніж інші типи, саме аудіальній частині спілкування (інтонації, тону голосу, паузам і т.п.). Щоб підтримувати контроль над відтворенням тональних аспектів свого мовлення, вони потребують повного дихання. Тому у них, як правило, більш широка грудна клітка у порівнянні з «візуалами», і вони використовують свої легені більш повно.

Нарешті, *«дискрети»* зазвичай кажуть затиснутим, твердим і монотонним голосом (використовуючи як і «візуали» верхню частину грудної клітки). Будова їхнього тіла схожа на кінестетиків, оскільки прийняття як бажану цифрову (логічну) репрезентативну систему зазвичай є засобом впоратися з якимись дуже важкими почуттями і переживаннями.

Таблиця 6.3 – Прояв репрезентативних систем

Патерни	«Візуал»	«Кінестетик»	«Аудіал»	«Дискрет»
Поза	Поза пряма, розправлена, голова і плечі підняті	Викривлена, зігнута, голова і плечі опущені	«Телефонна поза», голова нахилена в бік	Схрещені на грудях руки, пряма постава, піднята голова
«Типи» тіла і рухи	Як худий, так і повний, рухи скуті, судорожні	Пухкий, окрилений, м'який, рухи вільні, плавні	Нестійкий тип тіла, рухи то стиснуті, то вільні	М'яке, повне (далеко не завжди), рухи негнучкі
Дихання	Високе грудне	Низьке черевне	У повному обсязі	Обмежене
Тональність, швидкість і сила голосу	Високий, чистий, швидкий, гучний	Низький, неприродний, повільний, м'який	Мелодійний, ритмічний, мінливий	Монотонний, переривчастий, густий
Напрямок погляду за відношенням до оточуючих	Над оточуючими	Під оточуючими	Очі опущені	Дивиться над натовпом

Опис окорухових реакцій – відповідність напрямку погляду і режиму (характеру) роботи мозку людини (рис. 6.2):

1. Погляд вгору і вліво: візуальне конструювання (сконструйовані образи). Це візуальні образи або картини, які створюються індивідом. Сконструйовані образи зазвичай плоскі або недостатньо глибокі, а іноді і безбарвні.

2. Погляд вгору і направо: візуальні спогади. Це збережені візуальні образи або картини минулих подій, а також інших раніше відчутих візуальних подразників. Вони включають сни і конструйовані образи, які вже були випробувані. Ці образи зазвичай характеризуються як глибиною, так і рухом (як в кіно), а також і кольором.

3. Погляд горизонтально і вліво: аудіальне конструювання (сконструйоване мовлення). Цей патерн зазвичай пов'язаний з процесом створення мовлення. У цьому положенні людина «вкладає думки в слова», визначаючи те, що вона хоче сказати далі.

4. Погляд горизонтальний і направо: аудіальні спогади. Він включає в себе «алфавітну мелодику», літери, звуки реклами, номери телефонів, а також сленг і лайки. Подібний патерн зустрічається також, коли людина часто рухає очними яблуками при спогадах про почутий раніше слуховий образ, збережені у коротких, часто мелодійних або ритмічних патернах, чиє існування не усвідомлюється через часте повторення.

5. Погляд вниз і вліво: кінестетика (почуття). У цьому положенні очей людина отримує доступ як до емоцій і почуттів, що виникають, так і до збережених кінестетичних спогадів.

6. Погляд вниз і направо: внутрішнє мовлення (внутрішній діалог). Зазвичай пов'язаний з серйозними роздумами, коли цей процес супроводжується словами і звуками внутрішнього походження.

7. Розфокусовані очі: візуалізація. Цей патерн часто використовується під час розмови віч-на-віч між людьми, які спілкуються за правилом «подивися-послухай».

Ця схема «правильна» тільки для правшів. У лівшів все навпаки.

Рисунок 6.2 – Окорухові реакції

Експеримент: *Сядьте навпроти та уважно стежте за тим, куди підуть очі вашого партнера, після того, як ви задасте йому наведені нижче, груповані в блоки запитання, що застосовуються на тренінгах нейролінгвістичного програмування.*

Блок 1

1. Скільки дверей у вашому будинку?

2. Кого першим ви сьогодні побачили?

Очі вашого партнера, швидше за все, пішли *вгору і вправо*.

Блок 2

1. Уявіть, що парадні двері у вашому будинку пофарбовані в рожевий колір у зелений горошок.

2. Опишіть, яким ви себе побачили б на екрані телевізора.

Очі пішли *вгору і наліво*.

Блок 3

1. Почуйте, як дзвонить ваш будильник.

2. Почуйте, як шумлять хвилі.

Очі вашого партнера *по горизонталі йшли вправо*.

Блок 4

1. Яка ваша рука тепліша – ліва або права?

2. Яке у Вас буває відчуття, якщо прикусите язик?

Очі вашого партнера рухалися *вниз – вліво*.

Блок 5

1. Обговоріть самі з собою свої плани на майбутній місяць.

2. Згадайте про смішний епізод у минулому році.

Очі вашого партнера йшли *вниз і вправо* [7, с. 27–39].

Таким чином, нейролінгвістичне програмування вивчає суб'єктивний досвід, патерни, або «програми», створені взаємодією мозку (нейро), мови (лінгвістичне) і тіла, які обумовлюють поведінку [6, с. 528].

6.4. Комунікативна лінгвістика

Комунікативна лінгвістика є мовознавчою дисципліною нового типу, з метою вивчення свого об'єкта інтегрує різні галузі лінгвістики, а саме: теорію мовних актів, лінгвопрагматику, теорію мовленнєвої діяльності, дискурсологію, риторику і т.д.

Серед проблем комунікативної лінгвістики виділяють різні аспекти лінгвістики тексту, психолінгвістики, лінгвістичної семантики, стилістики, соціолінгвістики, лінгвокультурології, етнолінгвістики, когнітивної та комп'ютерної лінгвістики, теорії інформації і т.п. Комунікативна лінгвістика направлена на дослідження закономірностей, що складають чинники комунікативної діяльності, яка здійснюється на базі природної мови. *Об'єктом* комунікативної лінгвістики є мова, подана в реальних процесах комунікації (дискурсивна практика), а *предметом* – організація комунікативної ситуації у взаємодії її функціональних взаємно детермінованих модулів, одним з яких є вербальне повідомлення (текст).

Головними *завданнями* комунікативної лінгвістики є:

1) вивчення природи, типів і форм мовної комунікації;

2) виділення і опис мінімальних одиниць мовного спілкування (мовних актів, комунікативних актів, мовних подій, інтеракцій і т.д.), визначення їхньої ролі та ваги в комунікативному процесі;

3) характеристика типів комунікативної взаємодії та засобів комунікативного впливу;

4) обґрунтування зразків мовного спілкування: мовних жанрів і типів дискурсу;

5) аналіз тексту як знакової форми мовного спілкування і обґрунтування законів організації мовного коду в комунікації;

6) дослідження паравербальних і невербальних засобів мовної комунікації, їхніх функцій у комунікативній ситуації;

7) характеристика сфер мовної комунікації, їхньої взаємодії зі стилями спілкування;

8) вивчення психолінгвістичних аспектів породження, сприйняття і розуміння повідомлення та аналіз комунікативної компетенції у філогенезі та онтогенезі, її зв'язків з мовною компетенцією;

9) опис організації комунікативної ситуації, її складових та їхніх типів;

10) пошук оптимальних моделей комунікативної ситуації;

11) упорядкування методів аналізу мовної комунікації;

12) розробка рекомендацій щодо успішності, ефективності проведення комунікації, з її планування і контролю;

13) дослідження особливостей міжкультурної комунікації, її головних аспектів і понять;

14) розробка методик формування міжкультурної комунікативної компетенції тощо.

Головним поняттям комунікативної лінгвістики є ***комунікація*** як цілеспрямований процес інформаційного обміну між двома і більше сутностями за допомогою певної семіотичної системи.

Типи комунікації. ***Вербальна (мовна) комунікація*** – цілеспрямована лінгвопсихоментальна діяльність адресанта і адресата в процесі інформаційного обміну та впливу на співрозмовника (адресата) за допомогою знаків природної мови.

Невербальна комунікація є цілеспрямованим процесом інформаційного обміну, знаковими системами якого можуть бути біологічно доцільні поведінкові сигнали тварин, що визначають спільну адаптацію до навколишнього середовища; парамови жестів і міміки, математична і комп'ютерна символіка, мистецтво, гра, телепатичний зв'язок і т.д.

Проблема моделювання комунікативної ситуації. Вивчення роботи функціональних підсистем дискурсу і розуміння його природи може бути здійснено за допомогою моделювання – формалізованого і спрощеного подання комунікативної ситуації в сукупності всіх її складових і процесів, які опосередковують комунікативний акт.

Модель комунікативної ситуації (дискурсу) являє собою системну кореляцію певних складових, взаємодія яких сприяє інформаційному обміну і комунікативному впливу. Головними предметними компонентами моделей комунікації є комуніканти і текст (повідомлення).

Модель К. Бюлера. Першою комунікативною моделлю можна вважати модель знака К. Бюлера (рис. 6.3). К. Бюлер представляє власну модель повної конкретної мовної події в сукупності з життєвими обставинами, в яких вона зустрічається в певній мірі регулярно. Дослідник розміщує мову в центрі простору, в оточенні предметів і ситуацій, відправника та одержувача, які взаємодіють з мовою на підставі функцій репрезентації, експресії та апеляції відповідно.

Коло в середині символізує конкретне мовне явище, три змінні чинники покликані підняти його трьома різними способами в ранг знака. Три сторони наміченого трикутника символізують ці три фактори. Трикутник дещо менше, ніж коло (принцип абстрактної релевантності). Лінії символізують семантичні функції (складного) мовного знака. Це символ з точки зору співвіднесення з предметами і станом речей; це симптом на підставі своєї залежності від відправника, внутрішній стан якого він висловлює;

і сигнал в силу свого звернення до слухача, чиєю зовнішнією поведінкою або внутрішнім станом він керує так само, як і інші комунікативні знаки.

Рисунок 6.3 – Модель К. Бюлера

Технічна модель передачі інформації В. Вівера та К. Шеннона. Технічна схема передбачала наявність двох сигналів (від передавача і отриманого сигналу). Згідно зі схемою викривлення сигналу відбувалося в каналі через наявність технічного шуму (рис. 6.4).

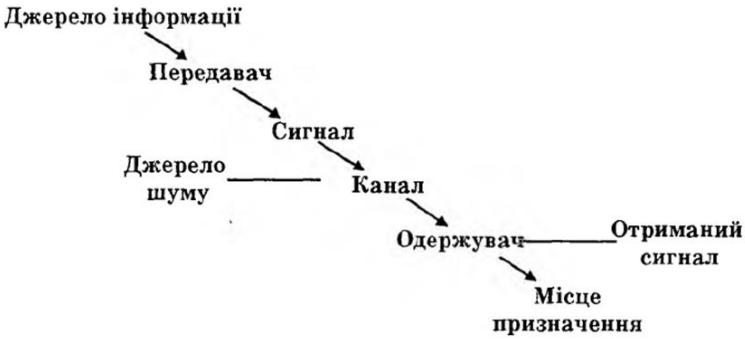

Рисунок 6.4 – Технічна модель передачі інформації
В. Вівера та К. Шеннона

Інформаційні моделі коммунікації. Технічна модель стала основою для створення інформаційних моделей комунікації. У. Еко модифікує модель В. Вівера та К. Шеннона, ввівши код як систему ймовірностей, що накладається на однакову ймовірність вихідної системи, забезпечуючи тим самим можливість комунікації (рис. 6.5).

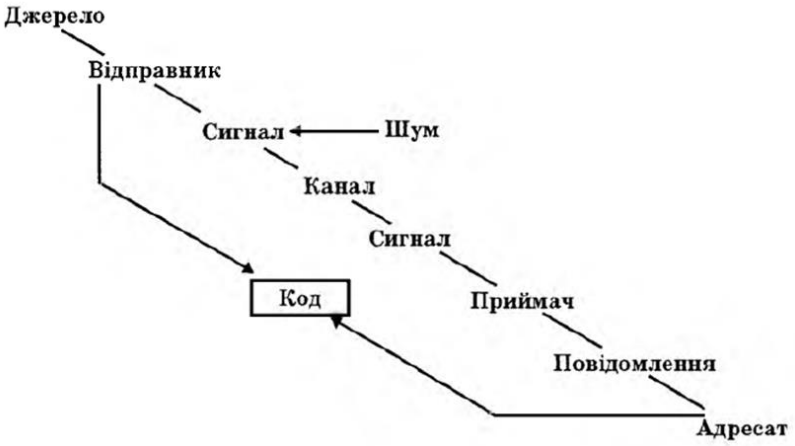

Рисунок 6.5 – Інформаційна модель комунікації У. Еко

Прагмалінгвістична модель І. Сусова. Дослідник представив замкнуту багатореляційну модель (рис. 6.6), в основі якої лежить трикутник, кути якого відповідають комунікантам і референтній ситуації, подібно до моделі К. Бюлера. Позитивними моментами цієї моделі є, по-перше, наявність одного висловлювання і двох сигніфікатів – смислів. Сигніфікат 1 є змістом, закладеним у висловленні адресантом-мовцем, враховуючи його інтенцію, а сигніфікат 2 є змістом, сприйнятим адресатом зважаючи на його інтерпретанту. По-друге, в моделі є дві прагматеми – інтенція та інтерпретанта, які корелюють з усіма складовими комунікативної ситуації і співвіднесені одна з одною. По-третє, всі складові моделі взаємопов'язані за принципом «діяльнісної системності» комунікативної ситуації.

Рисунок 6.6 – Прагмалінгвістична модель І. Сусова

Комунікативна взаємодія, її типи. Лінгвопрагматика. Процесуальною величиною комунікативної ситуації є комунікативна взаємодія (інтерактивність) як суб'єктно-об'єктно-суб'єктна діяльність учасників спілкування, спрямована на інформаційний обмін та вплив на свідомість або поведінку.

З огляду на ці параметри виділяються три типи комунікативної взаємодії. Першим типом є *комунікативна кооперація* (співробітництво), яка характеризується узгодженістю цілей, стратегічних програм комунікантів, симетричними відносинами між ними, балансом комунікативних статусів, ефективністю і оптимальністю спілкування.

Другим типом комунікативної взаємодії є *комунікативний конфлікт*, визначений неузгодженістю намірів, стратегічних програм комунікантів, асиметричними відносинами між ними, дисбалансом статусів особи, результатом чого переважно є припинення спілкування, небажання подальшого продовження комунікації.

Третім типом комунікативної взаємодії вважається *комунікативне суперництво* як неконфліктне здорове спілкування, що характеризується прагненням отримати інтелектуальну перевагу, здійснити свої наміри в диспуті, дискусії, підвищити власний статус особи.

Комунікативна взаємодія є предметом окремої галузі комунікативної лінгвістики – ***лінгвопрагматики***, що досліджує використання і функціонування мовних знаків у процесі комунікації у взаємозв'язку з інтерактивністю його суб'єктів (мовця і адресата), їхніми особливостями і самою ситуацією спілкування.

Сьогодні лінгвопрагматика інтегрує комунікативну лінгвістику з іншими галузями мовознавства, оскільки, з одного боку, прагматичне значення висловлювань незрозуміло за його інформаційним (семантичним) змістом, яке залежить від граматичної (синтаксичної) структури, з іншого боку, лінгвопрагматичне дослідження вимагає залучення даних психо-, етно- і соціолінгвістики, лінгвістики тексту, лінгвосеміотики, лінгвокультурології тощо.

Дослідники виділяють у лінгвопрагматиці три напрямки, орієнтовані на:

1) систематизацію прагматично заряджених мовних одиниць різних рівнів, вивчення їхньої взаємодії з семантикою і синтактикою;

2) дослідження інтерактивності комунікантів в процесах мовного спілкування;

3) моделювання когнітивних структур, що забезпечують інтерактивність дискурсів, зокрема, фреймів взаємодії, стратегічних програм.

У рамках першого напряму розглядаються проблеми пресупозицій, топікалізації, імплікації і т.д. Другий ґрунтується на принципах теорії мовленнєвих актів і направлений на емпіричні дослідження конкретних комунікативних ситуацій, встановлення їхньої типології, аналіз комунікативної взаємодії, способів оптимізації інтерактивності, аргументації, залежності мовлення від статусних і позиційних ролей комунікантів тощо. Лінгвопрагматика третього напряму застосовує теоретичний потенціал когнітивної науки та її галузі – когнітивної прагматики, вибудовуючи

основи для моделювання структур знань, що забезпечують стратегічне планування, хід і контроль комунікації, дію механізмів комунікативної компетенції, регуляцію процесів інтерактивності тощо.

Закони та правила комунікативної взаємодії. Перший постулат *кількості* визначає дозовану інформативність висловлювання:

1) зробіть своє висловлювання настільки інформативним, наскільки потрібно;

2) не робіть своє висловлювання більш інформативним, ніж потрібно.

Другий постулат *якості* вимагає правдивості, щирості та обізнаності мовця, тобто повідомлення повинно відповідати справжньому стану речей:

1) не говоріть того, що ви вважаєте хибним;

2) не говоріть того, для чого у вас немає достатніх підстав.

Третій постулат *відношення* визначає відповідність мовлення темі розмови:

1) будьте релевантними, говоріть по суті, відповідно до теми спілкування.

Четвертий постулат *манери мовлення* вимагає від мовця прозорості, зрозумілості, раціональності мовлення:

1) уникайте неясних висловлювань;

2) уникайте двозначності;

3) уникайте непотрібної надмірності;

4) будьте регламентовані.

На відміну від законів спілкування, *правила спілкування* є рекомендаціями щодо поведінки в тій чи іншій комунікативній ситуації:

1. *Перше правило* полягає в тому, що порядок аргументації повинен бути таким: сильний аргумент – середній – один найсильніший. Вважається, що слабкими аргументами не варто оперувати, а силу аргументів повинен визначити сам мовець.

2. *Друге правило* передбачає, що позитивній відповіді на важливе для адресанта питання повинні передувати дві позитивні відповіді адреса-

та на незначні і прості запитання. Такий результат запрограмований нейрофізіологічно, адже узгодження з кимось (відповідь «так») приводить до виділення ендорфінів – гормонів задоволення, а негативна відповідь супроводжується виділенням адреналіну, налаштовує людину на боротьбу та опір.

3. *Третє правило* ставить в залежність згоду людини від умов тільки почесної капітуляції, тобто, домагаючись переконання, не заганяйте співрозмовника в кут і дайте йому можливість зберегти свій статус особистості.

4. *Четверте правило* розглядає залежність переконливості аргументів від іміджу і статусу того, хто переконує. Авторитетність людини, її підтримка колективом нерідко посилює її аргументи при переконанні. Залежно від обставин найбільш переконливим є статус чоловіка, при інших умовах – жінки.

5. *П'яте правило* вимагає від того, хто стверджує, впевненої поведінки і неможливості зниження власного статусу особи.

6. *Шосте правило*, навпаки, забороняє зниження статусу співрозмовника, прояв неповаги, зневаги до нього.

7. *Сьоме правило* встановлює залежність переконання від особистого благородства і приємності мовця.

8. *Восьме правило* говорить: при переконанні треба починати не з того, що розділяє вас зі співрозмовником, а з того, що вас об'єднує, з чим ви погоджуєтеся.

9. *Дев'яте правило* ґрунтується на принципі емпатії і передбачає здатність до розуміння емоційно-психологічного стану співрозмовника в формі співпереживання. Краще розуміння партнера з комунікації оптимізує спілкування.

10. *Десяте правило* стосується уважного слухання, яке є запорукою переконливості.

11. *Одинадцяте правило* вимагає уникати конфліктогенних слів і висловів, які не сприятимуть переконанню, а тільки призведуть до припинення спілкування.

12. *Дванадцяте правило* стосується потреби в контролі за сприйняттям і розумінням співрозмовника, адже людині властиво ослабляти увагу і не завжди розуміти непрямі висловлювання.

13. *Тринадцяте правило* вимагає уваги до паравербальних засобів спілкування, їхньої відповідності вербальним, що в нейролінгвістичному програмуванні розглядається як конгруентність / інконгруентність. Невідповідність вербаліки і паравербаліки дозволить простежити за реакціями співрозмовника на аргументацію, а дотримання відповідності для тих, хто стверджує, значно посилить вагу його аргументів.

14. *Чотирнадцяте правило* направлено на демонстрацію того, що запропоноване вами задовольняє якусь потребу співрозмовника.

Виділяють серед цих правил активні, які посилюють позиції мовця (це 1, 2, 4, 7–10, 14), і пасивні, недотримання яких може послабити процес переконання або зруйнувати його (це 3, 5, 6, 11–13) [1, с. 550–616].

6.5. Соціолінгвістика

Соціолінгвістика, соціальна лінгвістика – міждисциплінарна наука, яка виникла на стику мовознавства, соціології, соціальної психології та етнографії і являє собою реалізацію на сучасному етапі лінгвістичних і соціолінгвістичних дослідницьких принципів та процедур соціологічного напрямку в мовознавстві. Соціолінгвістика вивчає широкий комплекс проблем, пов’язаних з соціальною природою мови, її суспільними функціями, механізмом впливу соціальних чинників на мову і тією роллю, яку відіграє мова в житті суспільства.

Мета і завдання загальносоціолінгвістичних досліджень носять багатовекторний характер:

1) вивчення мовної ситуації в державі з урахуванням якісних і кількісних даних щодо існуючих у ній мов і щодо суспільних функцій, які виконуються ними;

2) дослідження форм існування національних мов;

3) вивчення того, як використовують певну мову або будь-яку її форму існування представники різних соціальних груп відповідно до своїх етнічних, професійних, освітніх, вікових, гендерних та інших ознак;

4) вивчення загальних тенденцій мовного життя (мовної толерантності, конфліктності та ін.);

5) вивчення проблеми мови та ідеології;

6) вивчення соціолінгвістичної системи мовної поведінки;

7) дослідження проблеми білінгвізму і перемикання кодів;

8) дослідження проблеми мови і (само) ідентифікації;

9) вивчення варіаційної моделі мовних змін;

10) аналіз мовної політики і мовного будівництва держави для регулювання функцій, які виконуються мовами та її формами;

11) розробка функціональної класифікації в доповненні до існуючих генетичної та морфологічної класифікацій і виявлення їх співвідношень;

12) кінцева мета – вироблення соціолінгвістичних стратегій, що сприяло б гармонійному розвитку всіх мов.

Методики, прийоми і процедури аналізу. Методики соціолінгвістики являють собою комбінацію лінгвістичних і соціолінгвістичних процедур. За етапами дослідження їх можна поділити на:

1) методики і прийоми збору даних, матеріалу;

2) методики і прийоми аналізу;

3) методики і процедури верифікації.

1. *Методики збору соціолінгвістичних даних*:

– спостереження і включене спостереження;

– опитування інформантів через анкетування та інтерв'ювання;

– експерименти.

Ця методика, її прийоми та процедури отримали назву *активних методів*, оскільки дослідник сам визначає місце та учасників комунікації, форми проведення (спостереження, опитування, інтерв'ю, експеримент).

Основною перевагою активних (контактних) процедур і прийомів є живий контакт з інформантами, а недоліком – значна витрата часу і коштів,

а також обмежене охоплення мовних груп і ситуацій, тому активні методики слід поєднувати з пасивними (дистантними).

Другий етап – *методики і прийоми аналізу даних (пасивні).*

1. Аналіз документів:

а) матеріали перепису населення, статистичних збірників (наприклад, при вивченні білінгвізму в будь-якому регіоні доцільно спочатку встановити чисельність носіїв однієї та іншої мови, які проживають в даному районі, кількість шкіл, вищих і середніх навчальних закладів з викладанням кожною з цих мов, число радіо- і телепередач, періодичних видань, книг, що випускаються щорічно однією та іншою мовою, і т.д.);

б) персоніфіковані тексти (будь-які тексти в письмовій або усній формі, отримані в експериментальних умовах або спеціально відібрані, соціальні характеристики авторів, які відомі).

2. Обробка даних польових досліджень здійснюється, перш за все, за допомогою різновидів ***кореляційного аналізу***, суть якого полягає в співвіднесенні соціальних і мовних величин як незалежних і повністю або частково залежних змінних із залученням табличних даних, графіків залежностей і математичної статистики.

У *стратифікаційному різновиді* кореляційного аналізу виходять із співвідношення варіантів (форм) мови з певними спеціальними групами і виділяють професійні варіанти, жаргони, соціальні діалекти, варіанти ігрового характеру і т.д.

У *ситуативному різновиді* визначають варіанти мови залежно від конкретних соціальних ситуацій її вживання: обстановки і місця мовної події (університетська аудиторія, виступ у суді, офіційні переговори, сімейна бесіда і т.д.).

Комунікативна класифікація передбачає характеристику варіативності мови залежно від структури і форм мовлення: діалог чи монолог, усний чи письмовий, жанр, стиль мови і т.д.

Зазвичай ці кореляції описуються окремо за кожним виокремленим соціальним параметром, а потім піддаються змістовній інтерпретації та оформляються у будь-якому вигляді (таблиці, графіки тощо).

Останній етап соціолінгвістичних досліджень, пов'язаний з *верифікацією* отриманих даних на точність, достовірність, валідність (діагностична та прогностична сила, наприклад, анкети, опитування і т.д.), проводиться самим дослідником або експертами-фахівцями зазвичай з використанням імовірнісно-статистичного аналізу [6, с. 501–504].

Завдання

1. Навести графічне зображення польової організації концепту «лінгвістика».

2. Описати прийоми та методики психолінгвістики при:

– розпізнаванні брехні в мовленні;

– ідентифікації особистості за мовленням;

– мовленнєвому впливі і маніпуляції.

Література

1. Селіванова О. О. Сучасна лінгвістика: напрями та проблеми / О. О. Селіванова. – Полтава : Довкілля-К, 2008. – 712 с.

2. Попова З. Д. Когнитивная лингвистика / З. Д. Попова, И. А. Стернин. – АСТ : «Восток-Запад», 2007. – 226 с.

3. Кубрякова Е. С. Язык и знание: На пути получения знаний о языке: Части речи с когнитивной точки зрения. Роль языка в познании мира / Е. С. Кубрякова. – Москва : Рос. академия наук. Ин-т языкознания, 2004. – 560 с.

4. Глухов В. П. Психолингвистика. Теория речевой деятельности / В. П. Глухов, В. А. Ковшиков. – Москва : Астрель, 2007. – 238 с.

5. Белянин В. П. Психолингвистика / В. П. Белянин. – Москва : Флинта, 2004. – 232 с.

6. Комарова З. И. Методология, метод, методика и технология научных исследований в лингвистике / З. И. Комарова. – Екатеринбург : Изд-во УрФУ, 2012. – 818 с.

7. Ковалев С. В. Основы нейролингвистического программирования / С. В. Ковалев. – Москва : МПСИ, 2001. – 160 с.

8. Беликов В. И. Социолингвистика / В. И. Беликов, Л. П. Крысий. – Москва : Рос. гос. гуманит. ун-т, 2001. – 439 с.

9. Фрумкина Р. М. Психолингвистика / Р. М. Фрумкина. – Москва : Издательский центр «Академия», 2001. – 320 с.

10. Schmitt N. An Introduction to Applied Linguistics / N. Schmitt, M.P.H. Rodgers. – London: Routledge, 2019. – 392 p.

11. Bolshakov I. A. Computational Linguistics: Models, Resources, Applications / I. A. Bolshakov, A. Gelbukh. – Mexico City, 2004. – 186 p.

ДОДАТКИ

Додаток А
Теми рефератів

1. Проблема полісемії та омонімії.
2. Принципи побудови морфологічних аналізаторів.
3. Принципи побудови синтаксичних аналізаторів.
4. Аналіз систем машинного перекладу.
5. Статистичний підхід до розпізнавання мовлення.
6. Особливості роботи діалогових систем (чат-ботів).
7. Проблема перефразування у генеративній лінгвістиці.
8. Вирішення анафори та кореферентності.
9. Методи видобування колокацій.
10. Іменовані сутності та особливості їх видобування.
11. Аналіз тональності текстів як напрямок дослідження Natural Language Processing.
12. Методи машинного навчання в лінгвістиці.
13. Штучні мови. Лінгвопроектування.
14. Використання корпусу текстів для вирішення лінгвістичних завдань.
15. Лінгвістичне забезпечення інформаційних систем.
16. Створення та застосування онтологій.
17. Застосування частотних словників.
18. Мовна, наукова та національна картини світу.
19. Взаємозв'язок мови та мислення.
20. Методи судової психолінгвістики.
21. Стратегії та техніки нейролінгвістичного програмування.
22. Комунікативні тактики мовленнєвої діяльності.
23. Соціальна та територіальна диференціація мови.
24. Методи Social Network Analysis.
25. Концепції Інтернет-лінгвістики.

Додаток Б
Опис структури та правила оформлення реферату

Б1. Структура реферату

Оформлений реферат повинен містити: титульний аркуш, зміст, вступ, основну частину, а також висновки за результатами роботи.

У ***вступі*** розкривається актуальність теми, зазначаються мета і завдання дослідження. Обсяг вступу становить від 1 до 2 сторінок.

Основна частина реферату складається з 3–5 розділів, що розкривають тему роботи. У кінці кожного розділу формулюються висновки. Обсяг основної частини від 12 до 20 сторінок.

У ***висновках*** наводиться підсумок та виклад результатів, їх співвідношення з метою і завданнями, поставленими і сформульованими у вступі. Обсяг висновків становить 1–2 сторінки.

Список використаної літератури оформляється в порядку цитування у тексті (посилання на першоджерела) та становить не менше 10 пунктів.

Б2. Правила оформлення

Кожен розділ починається з нової сторінки:

– міжрядковий інтервал – 1,5;

– розмір шрифту Times New Roman – 14;

– параметри полів: ліве – 2,5 см, праве – 1,5 см, верхнє та нижнє – 2 см;

– абзаци починаються з відступу, що дорівнює 1 см;

– інтервал перед і після – 0 пт;

– вирівнювання тексту – по ширині, крім заголовків.

Заголовки розташовуються посередині рядка симетрично основному тексту роботи жирним шрифтом:

– відступ – 0 см;

– інтервал перед і після – 0 пт;

– вирівнювання тексту – по центру.

Додаток В

Зразок оформлення титульного аркуша реферату

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
«ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»

Кафедра інтелектуальних комп'ютерних систем

«Актуальні проблеми прикладної та теоретичної лінгвістики»

Реферат

«_____»

Виконав (ла):
студент групи _____
ПІБ

Перевірив (ла):
посада викладача
ПІБ

Харків, 20___

ЗМІСТ

ВСТУП 3
РОЗДІЛ 1
РОЛЬ ПРИКЛАДНОЇ ТА ТЕОРЕТИЧНОЇ ЛІНГВІСТИКИ В ІНФОРМАЦІЙНОМУ СУСПІЛЬСТВІ 5
1.1. Структура лінгвістики як науки 5
1.2. Сучасні напрямки розвитку прикладної лінгвістики 12
1.3. Лінгвістичні методи досліджень 13
Завдання 21
Література 22
РОЗДІЛ 2
МОДЕЛЮВАННЯ ПРИРОДНОЇ МОВИ ЗАСОБАМИ СТРУКТУРНОЇ ТА МАТЕМАТИЧНОЇ ЛІНГВІСТИКИ 23
2.1. Поняття лінгвістичної моделі 23
2.2. Формальні мови та граматики 27
2.3. Особливості моделювання природної мови 29
Завдання 33
Література 34
РОЗДІЛ 3
АКТУАЛЬНІ ПРОБЛЕМИ КОМП'ЮТЕРНОЇ ЛІНГВІСТИКИ 35
3.1. Прикладні завдання комп'ютерної лінгвістики 35
3.2. Машинний переклад 37
3.3. Автоматична генерація текстів 41
3.4. Створення діалогових систем 44
3.5. Розпізнавання та синтез мовлення 47
3.6. Класифікація та кластеризація 52
3.7. Витяг інформації 56
3.8. Аналіз тональності текстів 60
Завдання 61
Література 62

РОЗДІЛ 4
ПРИКЛАДНІ АСПЕКТИ КОРПУСНОЇ ЛІНГВІСТИКИ 63
4.1. Створення корпусу текстів 63
4.2. Розмітка корпусів 65
4.3. Застосування корпусних менеджерів 66
4.4. Використання корпусів 69
Завдання 71
Література 71
РОЗДІЛ 5
ЗАСТОСУВАННЯ СУЧАСНИХ ТЕХНОЛОГІЙ У ЛЕКСИКОГРАФІЇ 72
5.1. Класифікація словників 72
5.2. Макро- та мікроструктура словника 74
5.3. Комп'ютеризація словників 76
Завдання 79
Література 79
РОЗДІЛ 6
МІЖДИСЦИПЛІНАРНІ ДОСЛІДЖЕННЯ В ЛІНГВІСТИЦІ 80
6.1. Когнітивна лінгвістика 80
6.2. Психолінгвістика 88
6.3. Нейролінгвістика 99
6.4. Комунікативна лінгвістика 104
6.5. Соціолінгвістика 113
Завдання 116
Література 116
ДОДАТКИ. 118
Додаток А 118
Додаток Б 119
Додаток В 120

Навчальне видання

ПЕТРАСОВА Світлана Валентинівна
ХАЙРОВА Ніна Феліксівна

СУЧАСНІ ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ В ЛІНГВІСТИЦІ

Навчальний посібник
з курсу «Актуальні проблеми прикладної та теоретичної лінгвістики»
для студентів спеціальності «Прикладна та комп'ютерна лінгвістика»

Відповідальний за випуск проф. *Шаронова Н. В.*
Роботу до видання рекомендував проф. *Дмитрієнко В. Д.*

Редактор *Н. В. Верстюк*

План 2019 р., поз. 138

Підп. до друку 12.02.2020. Формат 60×84 1/16. Папір офсетний.
Riso-друк. Гарнітура Times New Roman. Ум. друк. арк. 5,9.
Наклад 50 прим. Зам. № 6/03/20. Ціна договірна.

Видавничий центр НТУ «ХПІ».
Свідоцтво про державну реєстрацію ДК № 5478 від 21.08.2017 р.
61002, Харків, вул. Кирпичова, 2

Видавець та виготовлювач: ФОП Панов А.М.
Свідоцтво серії ДК № 4847 від 06.02.2015 р.
м. Харків, вул. Жон Мироносиць, 10, оф. 6,
тел. +38(057)714-06-74, +38(050)976-32-87
copy@vlavke.com