Міністерство освіти і науки України
Миколаївський національний університет
імені В. О. Сухомлинського

ПРИКЛАДНЕ МОВОЗНАВСТВО (КВАНТИТАТИВНА І КОМП'ЮТЕРНА ЛІНГВІСТИКА, В ТОМУ ЧИСЛІ АВТОМАТИЧНА ОБРОБКА ПРИРОДНОЇ МОВИ)

(для студентів за освітньою програмою Філологія: Прикладна лінгвістика)

Навчально-методичні рекомендації

Миколаїв – 2017

УДК
ББК

Рекомендовано до друку Вченою радою філологічного факультету Миколаївського національного університету імені В.О.Сухомлинського (протокол № від)

Автор: к.філол. наук, старший викладач Мікрюкова К. О.

Рецензенти:

Доктор філологічних наук, професор кафедри

;

Кандидат філологічних наук, доцент кафедри мовно-літературної та художньо-естетичної освіти Миколаївського обласного інституту післядипломної педагогічної освіти – Д. В. Лук'яненко

Навчально-методичні рекомендації до курсу «Прикладне мовознавство (Квантитативна і комп'ютерна лінгвістика, в тому числі автоматична обробка природної мови)» (для студентів за освітньою програмою Філологія: Прикладна лінгвістика) : Навч.-метод. рекомендації. – Миколаїв : МНУ імені В. О. Сухомлинського, 2017. – 208 с.

У навчально-методичних рекомендаціях пропонується теоретичний матеріал до курсу «Квантитативна і комп'ютерна лінгвістика (в тому числі автоматична обробка природної мови)», що викладаються на третьому курсі за освітньою програмою Філологія: Прикладна лінгвістика, подаються завдання для формування практичних умінь та навичок, методичні матеріали для самостійної роботи, приклади модульних контрольних робіт та зразки підсумкових контрольних робіт, питання до іспиту, словник термінів курсу та список рекомендованої літератури.

ЗМІСТ

ВСТУП

Комп'ютери посідають важливе місце в житті сучасного суспільства. Комп'ютерна лінгвістика – це порівняно молода наука, що займається проблемами використання природної мови в системах автоматичної обробки інформації.

Курс «Квантитативна і комп'ютерна лінгвістика (в тому числі автоматична обробка природної мови)» передусім передбачає підготовку спеціалістів із автоматичної обробки текстів природною мовою для практичної роботи у сфері комп'ютерних лінгвістичних технологій. Студентам запропоновано низку теоретичних та практичних тем, присвячених дослідженню сучасного стану інтелектуальних технологій, методам автоматизованого аналізу текстів, системам машинного перекладу, мовам програмування для предметно-орієнтованих задач, комп'ютерним сервісам для роботи з корпусами текстів та базами даних. Навчально-методичні матеріали охоплюють десять основних тем курсу «Квантитативна та комп'ютерна лінгвістика як навчальна дисципліна», «Історія розвитку комп'ютерної лінгвістики», «Складові комп'ютерної лінгвістики. Спеціальні системи письма», «Аналітико-синтетичне опрацювання документів», «Проблеми створення штучного інтелекту», «Автоматичний синтаксичний аналіз», «Автоматичний семантичний аналіз», «Автоматичний морфологічний аналіз», «Машинний переклад як різновид інтелектуальних систем АОТ», «Комп'ютерні технології підготовки текстових документів». Крім цього, запропоновано методичні матеріали для самостійної роботи, приклади модульних контрольних робіт та зразки підсумкових контрольних робіт, питання до іспиту, словник термінів курсу та список рекомендованої літератури.

Мета навчальної дисципліни: вивчити комп'ютерні методи дослідження мови, сформувати у студентів уявлення про принципи обробки природної мови системами інформаційного пошуку, перекладу, управління, проектування. Ця дисципліна може служити підґрунтям для дослідницької роботи студентів.

Основними завданнями вивчення дисципліни є: дослідити опис основних типів комп'ютерних систем із лінгвістичним забезпеченням; вивчити основи комп'ютерних методів автоматичного аналізу текстів природною мовою; зрозуміти

принципи комп’ютерного аналізу текстів природною мовою; оволодіти основними методиками обробки текстів природною мовою та методами прикладної лінгвістики.

У результаті вивчення курсу студент оволодіває такими компетентностями:

І. Загальнопредметні: володіє розвиненою культурою мислення, умінням ясно й логічно висловлювати свої думки; володіє навичками наукової організації праці; розвиває навички самостійного опанування нових знань; уміє працювати з довідковою літературою, різнотипними словниками, електронними базами даних, системами інформаційного пошуку.

ІІ. Фахові: уміє застосовувати знання мови на практиці, користуватися мовними одиницями; знає одиниці мови та правила їх поєднання; критично оцінює набутий досвід із позицій останніх досягнень філологічної науки; володіє сучасною мовознавчою термінологією; володіє уміннями та навичками здійснювати мовленнєву діяльність, зумовлену комунікативною метою; обізнаний в основних напрямках комп’ютерної лінгвістики, механізмах комп’ютерної обробки інформації; засвоює основи комп’ютерних методів автоматичного аналізу текстів природною мовою.

Дисципліна «Квантитативна і комп’ютерна лінгвістика (в тому числі автоматична обробка природної мови)» базується на низці розділів курсів «Прикладна лінгвістика», «Сучасна українська літературна мова», «Теорія та практика перекладу», «Загальна та прикладна фонетика», «Експериментальні методи дослідження», «Інформатика», «Кібернетика», «Математичні основи гуманітарних знань». Успішне опанування дисципліни сприяє удосконаленню мовної та мовленнєвої компетентності студентів, позитивно впливає на їх загальний рівень освіченості та дозволяє у майбутньому бути конкурентоздатними на ринку праці.

ТЕМА 1
КВАНТИТАТИВНА ТА КОМП'ЮТЕРНА ЛІНГВІСТИКА ЯК НАВЧАЛЬНА ДИСЦИПЛІНА

План

1. Об'єкт, предмет, мета та завдання курсу квантитативна та комп'ютерна лінгвістика, в тому числі автоматична обробка природної мови.
2. Місце квантитативної та комп'ютерної лінгвістики у мовознавстві.
3. Місце квантитативної та комп'ютерної лінгвістики у кібернетиці.
4. Взаємозв'язок комп'ютерної лінгвістики з іншими науками.
5. Інформація. Типи інформації.

Теоретична частина

Проблема взаємозв'язку науки і практики в галузі лінгвістики розглядається з позиції вимог глобальної комп'ютеризації суспільства.

Прикладна лінгвістика другої половини XX ст. розглядає способи розв'язання практичних потреб мовлення не лише людиною, а й машиною. Цим зумовлено виникнення окремого напрямку прикладної лінгвістики – комп'ютерної лінгвістики.

Становлення комп'ютерної лінгвістики відбулося в Німеччині на базі комп'ютерної науки і структурно-математичної лінгвістики у 50-тих рр. XX ст.

Під терміном «комп'ютерна лінгвістика» розуміється широка область використання комп'ютерних інструментів – програм, технологій організації та обробки даних – для моделювання функціонування мови в тих чи інших умовах, ситуаціях, проблемних областях, а також сфера застосування комп'ютерних моделей мови не лише в лінгвістиці, а й у суміжних із нею дисциплінах. Сфера комп'ютерної лінгвістики охоплює практично все, що пов'язано з використанням комп'ютерів у мовознавстві. Термін «комп'ютерна лінгвістика» задає загальну орієнтацію на використання комп'ютерів для вирішення різноманітних наукових і практичних завдань, пов'язаних із мовою.

Комп'ютерна лінгвістика посідає проміжне місце між прикладною лінгвістикою та інформатикою.

Об'єктом комп'ютерної лінгвістики виступає мова в трьох формах свого існування:

- ✓ як мовна система (сукупність мовних одиниць із властивими їм формальними, змістовними властивостями);
- ✓ мовлення (різноманітні продукти реалізації мовної системи в певних умовах та ситуаціях комунікації);
- ✓ мовна діяльність (процес використання мовної системи і створення продуктів такого застосування мови в тих чи інших умовах комунікації).

Предмет дослідження комп'ютерної лінгвістики: ознаки будови, змісту та функціонування одиниць мовної системи, продуктів мовлення та мовної діяльності, що можуть служити засобом моделювання або опрацювання мовної інформації в комп'ютерній мережі.

Основні *проблеми* комп'ютерної лінгвістики:

- ✓ аналіз основних результатів наукових досліджень в галузі комп'ютерної лінгвістики;
- ✓ машинний переклад та проблеми, пов'язані з ним;
- ✓ моделювання спілкування;
- ✓ створення інформаційно-пошукових систем;
- ✓ автоматична обробка мовлення.

У 90-х р. XX ст. Ю. Городецький узагальнив *три класи проблем комп'ютерної лінгвістики*:

- ✓ 1 клас – пов'язаний із моделюванням спілкування (процес вираження емоцій, взаємодія вербаліки і невербаліки, розпізнавання комп'ютером різних повідомлень);
- ✓ 2 клас – охоплює проблему створення штучного інтелекту, розробка метамов та мов перекладу, перетворення мовної інформації іншою мовою);
- ✓ 3 клас – проблеми обробки письмового тексту (редагування текстів, реферування, анотування текстів, створення різних типів електронних словників).

Автоматична обробка природної мови – напрям комп'ютерної лінгвістики, який передбачає створення, перетворення й аналіз текстів із застосуванням природної або штучної мов, результатом чого може бути формування машинних фондів національних мов, автоматичних словників, термінологічних банків, комп'ютерних картотек, баз даних, комп'ютерних граматик тощо. Створення текстів відбувається у процесі автоматичного синтезу на підставі

семантичного, синтаксичного й лексемно-морфологічного представлення вихідної інформації. Перетворення текстів здійснюється при автоматизованому редагуванні – внесенні виправлень і доповнень, форматуванні – членуванні тексту, уведенні заголовка, нумерації сторінок і т. ін.; при лексикографічній обробці – уведенні щодо кожного слова словникової дефініції й іншої потрібної інформації або підготовці на базі текстів автоматизованих лексикографічних систем (лематизації); реферуванні – скороченні вихідного тексту відповідно до програми. Аналіз текстів передбачає членування тексту на одиниці, доступні комп'ютерній лексикографічній обробці, виокремлення ключових слів, персоналій, термінів. Аналіз і синтез текстів здійснюється на базі лінгвістичних процесорів – програмно-лінгвістичних комплексів багаторівневого типу, орієнтованих на граматичний, семантичний або когнітивний аналіз текстової інформації та діалог із користувачем. Процесори останнього покоління спрямовані на самонавчання шляхом гіпотез відносно явищ і категорій мови. Аналіз є складовою операцією машинного перекладу з однієї мови на іншу, а також перетворення текстів природною мовою на мови програмування, комп'ютерні кодові системи.

Квантитативна лінгвістика – це розділ мовознавства, спрямований на розроблення способів кількісного опису природних і штучних мов, статистичного підрахунку частотності різних мовних явищ у текстах. Квантитативна лінгвістика вважається складником більш широкої маргінальної дисципліни – математичної лінгвістики.

Одним із напрямів квантитативної лінгвістики є *лінгвостатистика*, завдання якої полягають у дослідженні частотності звуків, букв, слів, що стало підґрунтям для створення частотних словників різних мов; статистичному аналізі текстів із метою встановлення авторства, характеристики особливостей ідіостилю чи функціонального стилю; визначенні швидкості мовних змін; кількісному аналізі результатів психолінгвістичних експериментів.

Головний метод квантитативної лінгвістики – лінгвостатистичний експеримент, спрямований на отримання кількісних характеристик певних мовних явищ і встановлення достовірності статистичний результатів.

Комп'ютерна лінгвістика є дисципліною теоретичною, фундаментальною, і практичною (прикладною). Вона виявляє міжпредметні зв'язки як із дисциплінами прикладного спрямування, так і з власне теоретичними (фундаментальними) дисциплінами.

Лінгвістична дисципліна	*Розділ комп'ютерної лінгвістики*
Сучасна українська літературна мова (фонологія, акцентологія, морфологія)	Комп'ютерна граматика (вивчає основні засади побудови електронних повідомлень засобом ідентифікації яких в комп'ютері виступає червона або зелена стрічка).
Лексикографія	Комп'ютерна лексикографія (пропонує комп'ютерні версії традиційних словників, словниково зорієнтовані бази даних та ідеографічні показники).
Термінологія і термінографія	Комп'ютерна експертна система (комп'ютерна термінологія).
Лінгвістика тексту	Системи автоматичної обробки тексту.
Історія мови	Комп'ютерне шифрування або дешифрування.
Теорія і практика перекладу	Системи машинного перекладу

Назва «квантитативних лінгвістика» характеризує міждисциплінарний напрям у прикладних дослідженнях, в якому в якості основного інструменту вивчення мови та мовлення використовуються кількісні або статистичні методи аналізу.

Квантитативна лінгвістика тісно пов'язана з лексикологією. Звичайні тлумачні словники не вміщують у складі словникової статті інформацію про частоту використання тієї чи іншої лексеми. Це пов'язано із значною працемісткістю цього процесу. Для користувача словника така інформація може виявитися дуже важливою.

З теоретичної точки зору використання статистичних методів у мовознавстві дозволяє доповнити структурну модель мови імовірнісним компонентом, тобто створити структурно-імовірнісну модель, що має значний пояснювальний потенціал.

Кібернетика – наука про загальні принципи управління в різних системах. Основний об'єкт дослідження кібернетики –

кібернетичні системи.

Слово «кібернетика» (грец. «мистецтво керманича») вперше як термін для управління вжив давньогрецький філософ Платон. Кібернетика як наука виникла лише в 40-х роках XX століття. У 1948 році з'явилася перша праця, присвячена питанням кібернетики – «Кібернетика, або управління і зв'язок у тварині і машині». Її автор – відомий американський учений Норберт Вінер. Видатний французький вчений Ампер запропонував називати кібернетикою науку про управління людським суспільством.

Зв'язок комп'ютерної та квантитативної лінгвістики з кібернетикою виявляється у такій галузі мовознавства, як лінгвостатистика. Кількісні методи у лінгвістиці допомагають правильно організувати лінгвістичні спостереження, забезпечити надійність, точність, достовірність висновків.

Лінгвостатистика розглядається і як техніка обробки лінгвістичних даних, і як метод дослідження мови та мовлення, і як концепція, система ідей та уявлень про об'єкт лінгвістичної науки. І якщо техніка квантитативної лінгвістики використовується зараз у багатьох дослідженнях, то методика реалізується лише тоді, коли дослідник відчуває, що лише за допомогою кількісного підходу можна перевірити набуті знання про лінгвістичний об'єкт.

Можливість використання кількісних методів у мовознавстві базується на особливостях будови мови та мовлення. Мова – це система, яка складається з дискретних одиниць, що мають кількісні характеристики. Ці характеристики притаманні одиницям усіх рівнів. Мова має ймовірнісний характер: це код з імовірнісними обмеженнями.

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

1) дискретність одиниць;

2) масовість мовних одиниць;

3) повторюваність їх у висловлюваннях;

4) можливість вибору певного елемента з ряду однорідних. На мовлення впливають закони мови (закономірності будови одиниць мови, використання їх у мовленні), закони сполучуваності одиниць у мовленні, закони жанру, теми висловлювання, смаки автора, його психофізіологічний стан тощо. Дія цих факторів так переплітається, що інколи неможливо визначити результати їхнього впливу.

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

Комп'ютерна лінгвістика тісно взаємодіє з інформатикою. Проблеми інформатики в галузі розв'язання лінгвістичних задач висунули перед фахівцями з інформатики складне завдання: як за допомогою комп'ютерів обробляти інформаційні дані, подані у формі усних і письмових текстів природною мовою. Виникли технічні труднощі щодо збору, перетворення, збереження і передачі інформації, поданої у природному вигляді за допомогою сучасних комп'ютерних технологій, мова яких представлена у формалізованій формі.

Над створенням спеціальних пристроїв для автоматичного введення інформації в комп'ютер працюють інженери і фахівці з розпізнавання образів. Вже існують моделі електронних пристроїв, спроможні зчитувати машинний текст. Складнішим виявилось розв'язання проблеми, пов'язаної з читанням рукописного тексту. Але дослідження в цьому напрямку тривають. Наприклад, щоб підрахувати результати перепису населення, використовують електронні оптичні автомати «Бланк-6», здатні зчитувати рукописну цифрову інформацію.

Особливий інтерес становить розробка багато дикторських систем, здатних розпізнавати слова і фрази, вимовлені різними дикторами. Діалог «людина – машина» стане неефективним, якщо можливості людини будуть вивчені гірше, ніж можливості комп'ютера . Ця обставина підвищує важливість вивчення засобів автоматизованого діалогу, який стає комплексним із застосуванням засобів зорового, мовного, слухового і тактильного каналів.

Комп'ютерна лінгвістика нерозривно пов'язана з дослідженнями в галузі штучного інтелекту. *Штучний інтелект* – розділ комп'ютерології та інформатики (computer science), розробляє «розумні» комп'ютерні системи, тобто системи, які виявляють характеристики «розумності» в людській поведінці (розуміння мови, навчання, міркування, вирішення завдань тощо, а також етичні аспекти). Вивчення розуміння людської мови комп'ютером неможливе без заглиблення в теоретичні засади комп'ютерної лінгвістики.

Крім цього, знання у галузі комп'ютерної лінгвістики

застосовуються в біології, фізіології та медицині у вигляді комп'ютерної діагностики.

Із середини XX століття «інформація» стала загальнонауковим поняттям, але досі у науковій сфері воно залишається вкрай дискусійним. Загальноприйнятого визначення інформації не існує, і воно використовується головним чином на інтуїтивному рівні.

Інформація (від лат. informare – зображувати, повідомляти) – сукупність знань, образів, відчуттів, наявних у свідомості людини або штучному інтелекті, які поступають по різних каналах передачі, переробляються й використовуються у процесі життєдіяльності людини й роботі автоматичних комп'ютерних систем. У сучасній лінгвістиці інформація розглядається як така, що може бути маніфестована у знаковій формі природних мов, а також у паравербальних і невербальних засобах комунікації. Когнітивна й комп'ютерна лінгвістика оперують поняттям «інформація» в аспектах ментальних репрезентацій, структур представлення знань, процесів концептуалізації та категоризації, інформаційно-пошукових мов і систем. При автоматичній обробці природної мови у системах штучного інтелекту застосовуються також фонова інформація, яка є конвенційною та спільною для комунікантів і забезпечує зняття імплікативної невизначеності, двозначності, парадоксальності тощо.

Інформація має такі властивості:

1) цінність інформації – визначається корисністю та здатністю її забезпечити суб'єкта необхідними умовами для досягнення ним поставленої мети;

2) достовірність – здатність інформації об'єктивно відображати процеси та явища, що відбуваються в навколишньому світі. Як правило достовірною вважається насамперед інформація, яка несе у собі безпомилкові та істинні дані;

3) актуальність – здатність інформації відповідати вимогам сьогодення (поточного часу або певного часового періоду);

4) оперативність – властивість даних, яка полягає в тому, що час їхнього збору та переробки відповідає динаміці зміни ситуації;

5) ідентичність – властивість даних відповідати стану об'єкта;

6) суспільна природа – джерелом інформації є пізнавальна діяльність людей, суспільства.

7) мовна природа – інформація виражається за допомогою мови – знакової системи будь-якої природи, яка служить засобом спілкування, мислення, висловлювання думки. Мова може бути природною, а також штучною, створеною людьми з певною метою (наприклад, мова математичної символіки, інформаційно-пошукова, алгоритмічна тощо).

8) дискретність – одиницями інформації як засобами висловлювання є слова, речення, уривки тексту, а у плані змісту – поняття, висловлювання, гіпотези, теорії, закони тощо;

9) старіння – головною причиною старіння інформації є не сам час, а поява нової інформації.

За способом сприйняття інформація поділяється на: візуальну – сприймається органами зору; аудіальну – сприймається органами слуху; тактильну – сприймається тактильними рецепторами; нюхову – сприймається нюховими рецепторами; смакову – сприймається смаковими рецепторами.

За формою подання: текстову – що передається у вигляді символів, призначених позначати лексеми мови; числову – у вигляді цифр і знаків, що позначають математичні дії; графічну – у вигляді зображень, подій, предметів, графіків; звукову – усну або у вигляді запису передачі лексем мови аудіальним шляхом.

За призначенням існує: масова інформація – містить тривіальні відомості і оперує набором понять, зрозумілих більшій частині соціуму; спеціальна – містить специфічний набір понять; особиста – набір відомостей про яку-небудь особистість.

Питання для самоконтролю

1. Назвіть об’єкт, предмет, мета та завдання курсу квантитативна та комп’ютерна лінгвістика.
2. Визначте місце квантитативної та комп’ютерної лінгвістики у мовознавстві.
3. З’ясуйте місце квантитативної та комп’ютерної лінгвістики у кібернетиці.
4. Розкрийте зв’язок комп’ютерної лінгвістики з іншими науками.
5. Що таке інформація?

Завдання

1. Проаналізуйте основні типи інформації.
2. Укладіть словник термінів: *комп’ютерна лінгвістика, автоматична обробка природної мови, інформатика, інформація,*

інформаційне представлення тексту, штучна мова, штучний інтелект, кібернетика, тезаурус, квантитативна лінгвістика, лінгвостатистика, корпусна лінгвістика.

3. Створіть мультимедійну презентацію, що розкриватиме основні проблеми теми (20 слайдів).

4. Законспектуйте матеріал підручника Баранов А.Н. Введение в прикладную лингвистику. – М., 2001. – 345 с. (С.13–20).

ТЕМА 2
ІСТОРІЯ РОЗВИТКУ КОМП'ЮТЕРНОЇ ЛІНГВІСТИКИ

План

1. Етап виникнення комп'ютерних лінгвістичних систем.
2. Етап експериментальних комп'ютерних лінгвістичних систем.
3. Етап промислових комп'ютерних лінгвістичних систем колективного користування.
4. Етап промислових комп'ютерних лінгвістичних систем індивідуального користування.
5. Етап комп'ютерних лінгвістичних мереж.
6. Сучасний стан комп'ютерної лінгвістики в Україні.

Теоретична частина

Пристрої для автоматизації розрахунків існували здавна (рахівниці, арифмометри).

На початку XIX ст. із зростанням промислового виробництва зросла потреба в розрахунках, а також у керованих верстатах, що вимагали пристроїв для опрацювання інформації, наприклад, верстатів для ткання килимів.

У 1820–1830 рр. в Англії дослідник Ч. Бебідж створив першу в світі обчислювальну машину, яка давала змогу запам'ятовувати і вводити в ткацький верстат інформацію про візерунок килима. Збудована на дерев'яних «елементах», ця машина могла виконувати й математичні розрахунки. Цікаво, що вона складалася з основних компонентів сучасних комп'ютерів: зокрема, пристроїв уведення й виведення, пам'яті. Особливо корисним було те, що машина давала змогу змінювати програму.

Першим у світі програмістом, який писав програми для цієї машини, була Ада Лавлейс – донька лорда Дж. Байрона. Вже тоді вона геніально передбачила, що таку обчислювальну машину зможуть використовувати в багатьох галузях людської діяльності.

У 1890 р. в США під час перепису населення було

використано лічильно-перфораційні машини, в яких носієм інформації були перфокарти. До речі, після створення комп'ютерів інформацію в них ще довго вводили за допомогою саме таких перфокарт.

У 20-ті роки XX ст. були опубліковані дослідження американця Р. Хартлі про основи вимірювання кількості інформації, а в 30-ті – перші теоретичні роботи з кібернетики, в яких детально описувались ідеальні обчислювальні машини, названі за прізвищами їх творців – видатних математиків Тьюринга, Поста й Маркова.

Наприкінці 30-х на початку 40-х років у кількох країнах активно велися роботи зі створення електричних моделей обчислювальних машин, які мали ті ж основні компоненти, що й обчислювальна машина Ч. Бебіджа. Однією з причин створення таких машин була потреба в дешифруванні текстів у час Другої світової війни. Було створено кілька таких машин, проте вони виявилися недосконалими.

У 1943–1945 р. у Пенсільванському університеті інженери Дж. Еккерт і Х. Маклі створили ще одну таку машину, яка виявилася доволі вдалою і яку тому вважають першим комп'ютером. Ця машина отримала назву «ЕНІАК». Вона складалася з 18 тис. електронних ламп. На таких лампах пізніше працювали всі комп'ютери. Пізніше всі обчислювальні машини, в яких запам'ятовуючим елементом були лампи, віднесли до машин першого покоління.

Комп'ютери першого покоління виконували близько 10 тис. операцій за секунду. Мовами програмування в них були, як правило, машинні мови (асемблери) в цифрових кодах. Потім цифрові коди стали заміняти літерними позначеннями, внаслідок чого з'явилися спеціальні програми транслятори, які перекодовували літерні тексти програм у цифрові двійкові коди.

Відразу ж після Другої світової війни було опубліковано кілька теоретичних досліджень, які стали етапними в розвитку комп'ютерної техніки. Це книги Н. Вінера «Кібернетика» (1948 р.) і стаття К. Шеннона (у двох частинах) про теорію інформації (того ж 1948 р.).

У цей же час з'явилися перші комп'ютерні лінгвістичні системи, покликані до життя гострою соціальною потребою перекладу великої кількості науково-технічних текстів. Адже

середина XX ст. була часом науково-технічної революції, коли кількість науково-технічних публікацій зростала.

У 1946 р., відчуваючи актуальність задоволення цієї потреби, американці Р. Уівер і Дж. Бут сформулювали ідею комп'ютерного перекладу. Як результат у 1947 р. було написано першу програму для реалізації такого перекладу. Ця програма здійснювала так званий послівний переклад, тобто перекладала в текстах лише послідовно слово за словом, не виконуючи при цьому жодних інших процедур.

У 1952 р. в Масачусетському університеті відбулася перша конференція з комп'ютерного перекладу, а в 1954 р. в Джорджтауні публічно було продемонстровано можливості системи комп'ютерного перекладу. Словник цієї системи містив 500 слів, а під час експерименту було перекладено 400 речень (з англійської на російську мову). Для перекладу запропоновано речення: *It rains cats and dogs*, – яке, звичайно ж, було перекладено неправильно, оскільки є фразеологізмом. На початку 50 років проблеми комп'ютерного перекладу почали досліджуватися і в СРСР. Так, 1954 р. в Москві на з'їзді математиків дослідники Б. Мельчук і Х. Молошна виголосили доповідь про комп'ютерний переклад.

Паралельно з дослідженнями в галузі комп'ютерного перекладу тривали дослідження і в інших галузях комп'ютерної лінгвістики. У 1952 р. дослідники Р. Дадлі й Дж. Балашек (фірма Bell Telephone Laboratories) уперше оцифрували звуки мовлення, тобто перетворили аналогову інформацію в цифрову, ввели її в комп'ютер і вперше здійснили аналіз оцифрованих звуків. Результати дали змогу здійснити експериментальний комп'ютерний аналіз кількох слів (десяти цифр, вимовлених одним диктором). Для аналізу використовувався, зокрема, поріг у 900 Гц. Правильність розпізнавання становила 97%.

У 1958 р. ті ж дослідники провели аналіз звуків мови за десятьма діапазонами частот. Це дало змогу створити еталонні образи звуків мови.

Отож, за час з 40-х до початку 50-х років XX ст.:

– виникли комп'ютери як основний пристрій для реалізації комп'ютерних лінгвістичних систем;

– створено перші невеликі комп'ютерні словники, призначені для автоматичного перекладу текстів;

– у теорії комп'ютерного перекладу сформувалася

концепція двомовних відповідників, що забезпечувала послівний переклад;

– створено поодинокі експериментальні комп'ютерні лінгвістичні системи;

– виконано перші дослідження в галузі аналізу усного мовлення;

– сформульовано гіпотезу, про можливість математичного опису мови.

Усе це загалом дало підстави для гіперболізованого оптимізму щодо перспектив розвитку комп'ютерної лінгвістики.

У середині 50-х років для комп'ютерів замість ламп почали використовувати нову елементну базу – транзистори. Це дало змогу вмістити в одиниці об'єму більше запам'ятовуючих пристроїв, які, до того ж, споживали значно менше електроенергії. Ці комп'ютери назвали комп'ютерами другого покоління. Їхня швидкодія зросла до сотень тисяч операцій за секунду. Через дороговизну використовували як системи колективного користування, тобто за одним комп'ютером одночасно працювало багато користувачів.

Задачу кожного користувача операційна система ставила в чергу, далі виділяла однакові інтервали часу для кожної задачі й послідовно «по колу» виконувала задачі. Чим більше користувачів одночасно працювало в системі, тим менші інтервали часу виділяла операційна система на кожну задачу. Якщо ж один користувач завантажував на виконання не одну, а кілька задач, то робота ЕОМ ще більше сповільнювалася.

На цьому етапі для комп'ютерів були створені спеціальні програми, які суттєво спрощували їх роботу, – операційні системи. З'явилися також значно простіші, ніж асемблери, мови програмування – АЛГОЛ і ФОРТРАН, які стали називати *мовами високого рівня***.** Вся лексика, що використовувалася в програмному забезпеченні комп'ютерів, – лексика операційних систем і програмування – була реалізована на базі англійської мови (так склалося традиційно, хоча в СРСР існували й мови програмування на основі кирилиці). У 1963 р. використовували 30, в 1969 р. – 120, а в 1984 р. – вже кілька тисяч мов програмування.

У середині 50-х років Н. Хомський створив теорію формальних граматик, а також відповідний їй математичний апарат, який почали використовувати і в кібернетиці, і в лінгвістиці. Цю теорію відразу ж почали широко досліджувати,

проте, як засвідчила практика, її застосування виявилося ефективним тільки в трансляторах, що опрацьовували тексти програм (використання в лінгвістиці виявилося неефективним).

У цей час у лінгвістичних дослідженнях почали широко застосовувати теорію інформації К. Шеннона і, як показав час, не завжди обґрунтовано.

В СРСР на початку 60-х років дослідник Д. Панов висунув ідею автоматизації редагування, розглядаючи редагування як переклад з мови не зовсім правильної на мову правильну.

У 1956 р. в Італії було здійснено першу спробу автоматизації лексикографічного опрацювання текстів, написаних літерами латинської, грецької та кириличної абеток.

У цей період були створені комп'ютерні лінгвістичні системи переносу слів (Франція), коректури тексту (СРСР), досконалі системи шифрування/дешифрування текстів (дипломатичні, розвідувальні й силові структури багатьох держав).

На той час уже з'ясувалося, що здійснювати переклад на основі теорії двомовних відповідників у промисловому режимі недоцільно (переклади були настільки недосконалими, що зрозуміти їх було вкрай важко). Тому з'явилася нова ідея: створити штучну мову-посередник, на яку перекладали б природномовні тексти, а потім з мови-посередника – на потрібні природні мови. Це могло дати суттєвий виграш: кількість мов, потрібних для функціонування систем комп'ютерного перекладу, різко скорочувалася. Цілі наукові колективи почали працювати над створенням штучних мов-посередників, використовуючи для цього як штучні, так і природні мови. Проте, як з'ясувалося після багаторічної виснажливої роботи, створити таку штучну мову-посередник неможливо. Якби таку мову й вдалося створити, то, фактично, вона виявилася б природною.

На шляху перекладу без мови-посередника дедалі частіше почали з'являтися нові нездоланні труднощі на зразок омонімії (наприклад, як в українському слові *замок***,** англійському *set* тощо). Яке зі значень цих слів потрібно обрати для перекладу в кожному конкретному випадку, програмно з'ясувати було неможливо, а тому обирали найпоширеніше, що часто спричиняло помилки.

Паралельно тривали роботи з опрацювання усних текстів. У 1959–1960 рр. дослідникам Дж. Дінесу й Х. Метьюзу для кращого зіставлення з еталонними образами вдалося нормалізовувати

(«розтягувати» й «стискати» в часі) слова усної мови.

У 1962 р. на Всесвітній виставці в Сіетлі американська фірма IBM продемонструвала розпізнавач усного мовлення, який вміщувався у валізі.

Отож, за час зі середини 50-х до середини 60-х років XX ст.:

– на комп'ютерах було укладено кілька частотних словників для основних мов світу (найчастіше словоформ, без лематизації);

– для деяких мов було реалізовано автоматичний морфологічний аналіз тексту, хоч і ще не достатньо ефективний (поділ слів на морфеми й приписування словам граматичних категорій);

– деяких успіхів було досягнуто в синтаксичному аналізі, хоча основні труднощі (наприклад, при синтаксичній омонімії в реченні *Кабінет Міністрів попередив Парламент)* так і не вдалося подолати;

– усі комп'ютерні лінгвістичні системи були експериментальними;

– усі лінгвістичні задачі виявилися в сотні чи в тисячі разів складнішими, ніж передбачалося;

– різко скоротилося державне фінансування наукових робіт і кількісний склад персоналу цього напряму.

У середині 60-х років інженерні працівники знову вдосконалили елементну базу комп'ютерів: транзистори було замінено на інтегральні схеми (кожна така схема містила одночасно десятки чи навіть сотні транзисторів). Такі комп'ютери назвали комп'ютерами третього покоління. Їх швидкодія сягала мільйона операцій за секунду. Як і комп'ютери другого покоління, ці комп'ютери були системами колективного користування.

У складі операційних систем почали створювати програми для набирання текстів, які давали змогу вводити їх у пам'ять комп'ютерів відразу з перфокарт, перфострічок чи з клавіатур, при цьому текст, уведений з клавіатури, відтворювався на екрані електронно-променевої трубки – дисплея. Ці програми англійською стали називати *text editor* (у перекладі – текстовий редактор, чи виправляч текстів, текстовий виправляч*).*

У лінгвістичних дослідженнях почали широко використовувати різні математичні апарати, зокрема математичну логіку, теорію графів, формальні граматики, статистику, теорію

інформації тощо. Це давало змогу створювати математичні моделі одиниць мови, мовлення й текстів. Було також більш детально вивчено закон Ципфа. В останні роки цього періоду виникла теорія нечітких множин, яка дала змогу більш адекватно моделювати явища мови.

На цьому етапі певних успіхів досягла комп'ютерна лексикографія. Так, 1968 р. розпочалося укладання на комп'ютері логотеки шведської мови, в якій для кожного слова було передбачено поля, що містили таку інформацію: правописну, фонетичну, морфологічну, семантичну, синтаксичну, фразеологічну, стилістичну, етимологічну. У1969 р. – словника італійської мови (було укладено словник на 106 тис. слів); 1972 р. – на основі опрацювання текстів довжиною 90 млн. слововживань – 15-томного словника французької мови на 175 тис.

Враховуючи, що опрацювання зв'язних текстів на попередньому етапі виявилося непосильним завданням, мовознавці в експериментальних дослідженнях звернулися до опрацювання незв'язних текстів. Тут чудовим полігоном для постановки лінгвістичних задач виявилися інформаційно-пошукові системи, які давали змогу опрацьовувати як окремі слова, так і цілі словосполучення. У результаті було укладено інформаційно-пошукові тезауруси, проведено низку досліджень повноти й точності інформаційного пошуку.

Паралельно досліджувалися й створювалися системи індексування текстів, які також базувалися на опрацюванні незв'язних текстів – окремих слів і словосполучень. Розпочалися роботи і з автоматичного реферування з метою використання його в інформаційно-пошукових системах. Наслідком досліджень стало створення цілої низки інформаційно-пошукових систем, які функціонували на основі ключових слів чи дескрипторів, систем індексування (наприклад, системи індексування Ланкастера), перших експериментальних систем реферування, бібліотечних систем (систем бібліографування).

Попри скорочення фінансування не припинялося опрацювання зв'язних текстів, велося детальне експериментальне розв'язання невеликих лінгвістичних задач (створення моделей перекладу окремих типів синтаксичних конструкцій, синтезування потрібних варіантів словоформ тощо).

У ділянці створення систем комп'ютерного перекладу

відбулося переосмислення його суті. Так, було запропоновано й визнано доцільним: а) здійснювати переклад не в автоматичному, а в автоматизованому режимі, поступово навчаючи систему правильно перекладати найуживаніші слова й словосполучення; б) знизити вимоги до перекладеного тексту, тобто допускати синтаксичну непов'язаність слів і словосполучень; в) визнати необхідним проводити після комп'ютерного перекладу ручне після редагування (постредагування). Внаслідок великих фінансових вкладень й інтелектуальних потуг з'явилися перші промислові системи комп'ютерного перекладу, серед яких – SYSTRAN, «General Motors».

У СРСР роботи зі створення систем комп'ютерного перекладу велися в Санкт-Петербурзі під керівництвом Р. Г. Піотровського (англійсько-російський і російсько-англійський переклади).

Відбулося багато міжнародних конференцій з комп'ютерної лінгвістики, що дали лінгвістам змогу узагальнити результати, отримані науковими колективами.

Отже, від середини 60-х до початку 80-х років XX ст.:

– досліджувалися методи аналітико-синтетичного опрацювання текстів – індексування, пошуку, реферування;

– залишилися труднощі в синтаксичному аналізі зв'язного тексту;

– дослідники впритул підійшли до необхідності семантичного аналізу тексту;

– нагромаджено великий практичний досвід конструювання комп'ютерних лінгвістичних систем;

– підготовлено цілу низку комп'ютерних словників, які принципово відрізняються від аналогічних паперових, і призначені для автоматичного аналізу й синтезу текстів;

– вагомими були здобутки аналізу й синтезу усного мовлення.

На початку 80-х років елементна база комп'ютерної техніки знову була вдосконалена. На зміну інтегральним схемам, які назвали малими, прийшли надвеликі. Одна така мікросхема містила вже тисячі, десятки й сотні тисяч транзисторів. Це визначило напрям розвитку комп'ютерної техніки цього етапу – мініатюризація. У результаті замість великих, середніх і малих комп'ютерів, які займали цілі зали, було сконструйовано так звані персональні комп'ютери. Їх було названо комп'ютерами четвертого

покоління. Вони поміщалися на столі, а деякі з них можна було, як портфель, носити зі собою. Швидкодія таких комп'ютерів сягала десятків і сотень мільйонів операцій за секунду. За обсягом пам'яті, як тимчасової, так і постійної, вони не тільки не поступалися своїм попередникам.

Перший комп'ютер цього типу 1976 р. зібрали С. Возняк та С. Джобс, які пізніше створили комп'ютерну фірму Apple. Проте провідну роль у масовому випуску персональних комп'ютерів (ПК) відіграла не фірма Apple, а фірма IBM, яка 1981 р. випустила дешевшу модель персонального комп'ютера – IBM XT. Трохи пізніше процесори до персональних комп'ютерів почали випускати інші американські фірми – Intel, Motorola, AMD. Відтоді американська комп'ютерна індустрія стала законодавцем тенденцій розвитку, перетворившись практично в монополіста випуску комп'ютерної техніки, зокрема їх «ядра» – процесорів.

У програмному забезпеченні на одне з чільних місць виступає проблема уніфікації роботи операційних систем. Адже на попередніх етапах кожна операційна система мала свою унікальну систему команд. Визначну роль у цих процесах почала відігравати американська фірма Microsoft, яка для IBM-сумісних комп'ютерів створювала найбільш прості й надійні в користуванні операційні системи. Ці програми мали графічні (у вигляді малюнків) засоби спілкування з користувачами. Задля справедливості слід зауважити, що й тут пріоритет мали програмісти фірми Apple, які першими створили графічні операційні системи.

На цьому етапі продовжувалось укладання словників різних мов. Так, 1984 р. в Німеччині на основі обстеження текстів довжиною 50 млн. слововживань укладено словник на 90 тис. слів. У 1986 р. в комп'ютер було введено 13-томний Оксфордський словник англійської мови, а також 4-томне доповнення до нього. Того ж 1986 р. в університеті японського міста Кіото за допомогою комп'ютерної техніки було укладено словник японської мови. Аналогічні роботи проведено і в деяких інших країнах (Чехія, Польща, Угорщина). Крім того, у 80-ті роки реалізовано міжнародний проект LDOCE, який передбачав укладання словника англійської мови для тих, хто її вивчає. Результатом стало укладання словника на 40 тис. слів, які мають 76 тис. дефініцій (в цих дефініціях використано лише 2 тис. слів).

Серед експериментальних лінгвістичних досліджень цього

часу треба виділити роботу американського дослідника Т. Вінограда, який створив робота, що розумів команди, подані природною мовою, і виконував накази (переносив предмети). Це була одна з перших комп'ютерних систем розуміння текстів (фактично, окремих речень). Її створення дало поштовх формуванню теорії семантичних сіток, фреймів і сценаріїв. У цей час з'являються перші монографічні дослідження й підручники, а також перші періодичні видання з комп'ютерної лінгвістики.

Особливо важливо, що на цьому етапі розвитку комп'ютерної лінгвістики з'являється ринок промислових, хай і не завжди достатньо досконалих, комп'ютерних лінгвістичних систем.

Отже, з початку 80-х до початку 90-х років XX ст.:

– комп'ютерні лінгвістичні системи із систем колективного користування перетворилися на системи індивідуального користування;

– у розвинених країнах комп'ютерні лінгвістичні системи почали входити в повсякденне життя людей;

– з'явився ринок комп'ютерних лінгвістичних систем, на якому виникла конкуренція за ринки збуту;

– на перехрещенні кібернетики й мовознавства почала формуватися нова галузь мовознавства – комп'ютерна лінгвістика;

– в університетах починається підготовка спеціалістів з комп'ютерної лінгвістики.

На цьому етапі *комп'ютерних лінгвістичних мереж* суттєвих змін в елементній базі комп'ютерів не відбулося. Комп'ютери продовжували працювати на тих самих надвеликих інтегральних схемах, проте кількість транзисторів на одній такій мікросхемі суттєво зросла й досягла 20 млн. при розмірах у половину сірникової коробки. Крім IBM-сумісних комп'ютерів, випускалися комп'ютери інших моделей (Apple на основі процесорів Motorola, великих комп'ютерів на основі процесорів Alpha та ін.), але їх було значно менше. Швидкодія комп'ютерів сягнула сотні й тисячі мільйонів операцій за секунду (1, 2, 3 ГГц). На початку XXI ст. на ринку з'явилися кишенькові комп'ютери.

Особливо важливим було те, що, починаючи з цього етапу, комп'ютери стали мультимедійними, тобто отримали можливість цифрового опрацювання звука й відео (графічної динамічної інформації).

Висловлюючись образно, починаючи з цього періоду,

комп’ютери отримали «органи» зору, слуху й мовлення, а, отже, стали більше схожі на людину.

У програмному забезпеченні набули поширення альтернативні операційні системи – Unix та Linux, призначені для організації роботи та функціонування комп’ютерних мереж.

Проте вирішальним на цьому етапі було не стільки вдосконалення комп’ютерів, скільки розширення функціонування комп’ютерних мереж, зокрема виникнення глобальної комп’ютерної мережі – Інтернету. Але водночас постала й зовсім нова лінгвістична проблема, пов’язана з необхідністю єдиного шрифтового забезпечення, яке б давало змогу спілкуватися носіям усіх писемностей Земної кулі – як літерних, так й ієрогліфічних (наявне шрифтове забезпечення такої можливості не давало).

У створенні комп’ютерних лінгвістичних систем провідну роль почав відігравати Інтернет. У відповідь на потреби Інтернету з’явилися шрифти нового типу – *UNICODE*, кожен комплект яких містить приблизно по 65 тис. знаків, зокрема літери, лігатури й ієрогліфи для всіх основних мов світу.

Поширення набули нові типи словників, так звані електронні словники, а також комп’ютерні – на оптичних дисках (як одномовні, так і перекладні). Електронні словники почали виготовляти у формі спрощених кишенькових комп’ютерів, у пам’ять яких записували лексикографічну інформацію. У найдосконаліших з них кількість слів сягала кількох сотень тисяч, а кількість мов – до десяти-двадцяти. Комп’ютерні словники використовували на стаціонарних і портативних комп’ютерах.

З’явилися комп’ютерні лінгвістичні системи, що забезпечують можливість керування роботою систем з голосу, в тому числі електронні «секретарі» – програми, що писемно фіксували текст, який диктує людина.

Загалом, на цьому етапі зросла якість функціонування всіх видів комп’ютерних лінгвістичних систем. На ринку інформаційних технологій продаж комп’ютерних лінгвістичних систем почав давати великі прибутки.

Отож, із середини 90-х років XX ст. до початку XXI ст.:

– укладено шрифти UNICODE, що містять абетки й ієрогліфи більшості мов світу;

– значна частина комп’ютерних лінгвістичних систем через канали зв’язку Інтернету стала доступною користувачам (в

основному громадянам розвинених країн);

– з'явилися перші комп'ютерні лінгвістичні системи, які здійснювали частковий семантичний аналіз у текстах вузької тематики;

– зросла якість функціонування всіх видів комп'ютерних лінгвістичних систем;

– постала проблема спряження (тобто взаємоузгодження) й стандартизації даних у комп'ютерних лінгвістичних системах, зокрема, лексикографічних (словникових);

– зросла кількість працівників, зайнятих у сфері комп'ютерної лінгвістики;

– у вищих навчальних закладах розпочалася підготовка фахівців для галузі комп'ютерної лінгвістики.

З початку XX ст. Україна перебувала в складі СРСР. Правлячий в СРСР тоталітарний режим втручався в розвиток науки. Внаслідок таких втручань після Другої світової війни, коли в країнах Європи й США бурхливо почала розвиватися кібернетика, в СРСР її було визнано лженаукою. Це загальмувало її розвиток приблизно на десятиліття.

Попри це перший в СРСР комп'ютер – Малу електронну лічильну машину (МЕЛМ) – було створено в Києві в 1951–1953 рр. Очолював роботу академік С. Лебедєв (Інститут електротехніки Академії наук УРСР). Трохи пізніше в СРСР було створено цілу низку оригінальних комп'ютерів й операційних систем, які за своїми експлуатаційними характеристиками не поступалися західним аналогам.

У 60-ті роки в Інституті кібернетики АН УРСР, який працював під керівництвом академіка В. Глушкова, була сформульована можливість створення й розвитку малих комп'ютерів. Проте від цієї продуктивної ідеї через насаджений в СРСР культ гігантоманії помилково відмовилися, що в майбутньому суттєво загальмувало перехід на персональні комп'ютери. Водночас через відставання в розвитку елементної бази, всі комп'ютери, які створювали в СРСР, почали копіювати з західних (американських, французьких).

У 60-70-х роках в Україні вперше у світі було випущено двотомну «Енциклопедію кібернетики».

Центром розвитку комп'ютерної техніки в Україні став Київ, де в 70-80-х роках випускали середні й малі комп'ютери типу СМ (СМ-4, СМ-1420, СМ-1800 тощо).

Дослідження в комп'ютерної лінгвістики в Україні проводив відділ структурно-математичної лінгвістики Інституту мовознавства ім. О. Потебні. Зокрема, цей відділ уклав перший в Україні двотомний «Частотний словник сучасної української художньої прози». У цьому ж відділі створено програми лематизації словоформ української мови тощо. Проте загалом через консерватизм керівництва, відсутність комп'ютерної техніки дослідження української мови в руслі комп'ютерної лінгвістики велося до 90-х років у недостатніх обсягах. Саме тому не було створено програмного забезпечення (програм перевірки орфографічної правильності текстів, систем комп'ютерного перекладу, комп'ютерних словників тощо), яке дало б змогу опрацьовувати україномовні тексти.

Після здобуття Україною незалежності, 1993 р. в НАН було створено Український мовно-інформаційний фонд (фактично – науково-дослідний інститут). Його завданням став випуск 2002 р. оптичного диска «Словники України», на якому в формі інтегрованої бази даних записано п'ять словників української мови: орфографічний парадигматичний (на 152 тис. слів), орфоепічний, синонімічний, фразеологічний та антонімічний. До кожного слова подано приклади використання в текстах класичної української літератури – художньої чи науково-технічної. Заплановано також доповнення цього словника тлумачним й омонімічним словниками, а також представлення всієї інтегрованої лексикографічної бази даних в Інтернеті.

У 1997 р. почав виходити електронний «Український журнал комп'ютерної лінгвістики».

Наприкінці 90-х років створено систему «РУТА-ПЛАЙ», яка давала змогу перевіряти орфографічну правильність україномовних текстів і здійснювати українсько-російський та російсько-український переклад. Пізніше система була доповнена функціями контролю стилістичної та пунктуаційної правильності україномовних текстів.

Зараз в Україні в ділянці комп'ютерної лінгвістики працюють такі відомі науковці, як Людмила Алексієнко, Наталія Дарчук, Євгенія Карпіловська, Наталія Бардіна, Валентина Перебийніс, Тетяна Грязнухіна, Анатолій Анісімов, Тарас Вінцюк та інші.

Завдання української комп'ютерної лінгвістики полягають у тому, щоби створити україномовні комп'ютерні лінгвістичні

системи. Без них Україна не зможе увійти в міжнародне інформаційне співтовариство.

Питання для самоконтролю

1. Які основні досягнення здійснено на етапі виникнення комп'ютерних лінгвістичних систем?
2. Назвіть основні досягнення етапу експериментальних комп'ютерних лінгвістичних систем?
3. Чим відзначився етап промислових комп'ютерних лінгвістичних систем колективного користування?
4. Що нового винайдено на етапі промислових комп'ютерних лінгвістичних систем індивідуального користування?
5. В чому специфіка етапу комп'ютерних лінгвістичних мереж?

Завдання

1. Проаналізуйте сучасний стан комп'ютерної лінгвістики в Україні.
2. Підготуйте реферат на одну із тем: мови країн Близького Сходу (на вибір – арабська, перська, турецька); мови країн Далекого Сходу (на вибір – китайська, японська, корейська); мови країн Африканського континенту (на вибір – суахілі, банту).
3. Законспектуйте матеріал підручника Баранов А. Н. Введение в прикладную лингвистику. – М., 2001. – 345 с. (С.20–38).

ТЕМА 3
СКЛАДОВІ КОМП'ЮТЕРНОЇ ЛІНГВІСТИКИ. СПЕЦІАЛЬНІ СИСТЕМИ ПИСЬМА

План

1. Прикладна фонетика: мовний звук; частота звука; інтенсивність; тривалість; спектральний склад звуків; інструменти аналізу звуків.
2. Традиційні системи письма: поняття про письмо; види систем письма; система письма; критерії оцінювання графіки; пристрої для традиційного письма.
3. Транскрибування: звукове транскрибування, інтонаційне транскрибування. Застосування транскрибування.
4. Транслітерування.
5. Стенографування.
6. Системи письма для незрячих.

7. Мова жестів.

8. Криптографування. Азбука Морзе. Цифрові системи письма.

Теоретична частина

Серед звуків, які генерує чи сприймає людина, виділяють немовні й мовні (лінгвальні). *Мовний звук* – це звук лише з певними акустичними характеристиками, які утворюють його значення серед інших лінгвістичних звуків. На відміну від слів, які позначають фрагменти світу, мовні звуки таких фрагментів не позначають. Звичайно, тут ми не враховуємо ситуації, за якої, наприклад, двоє людей попередньо домовляються: «Якщо я гукну «А», то це означатиме, що наші друзі вже почали святкування Нового року».

Нагадаємо, що усне мовлення стосовно писемної мови є первинним. Порівняйте характеристики цих двох видів мовлення: швидкість усного мовлення (без пауз) – 2,0-3,6 слова/с; швидкість письма – 0,4 слова/с; швидкість друкування: непрофесійного – 0,2-0,4 слова/с; професійного – 1,6-2,5 слова/с.

Як бачимо, усне мовлення в кілька разів швидше за писемне. Це свідчить про неабияку актуальність дослідження усного мовлення, зокрема, з метою створення комп’ютерних лінгвістичних систем.

Теоретична фонетика, як правило, основну увагу приділяла дослідженню опозицій, в які вступають мовні звуки (наприклад: голосні – приголосні, глухі – дзвінкі). Прикладна фонетика основну увагу приділяє дослідженню звуків у інших аспектах: артикуляторному (артикуляція анатомічних органів гортані людини під час утворення звуків); фізіологічному (психофізіологічні й нейролінгвістичні механізми сприйняття і розпізнавання звуків); акустичному (фізичні характеристики звуків, згенерованих людиною, та можливості сприйняття органів слуху).

У прикладній фонетиці використовують цілу низку пристроїв, що дають змогу вивчати органи мовлення людини, і різні методи дослідження: палатографію (встановлюються місця торкання язиком піднебіння під час утворення звуку); тензопалатографію (для вимірювання сили артикуляції); рентгенографію та кінорентгенографію (для спостереження за позицією органів мовлення та їх переміщеннями); осцилографію (для визначення тривалості, частоти й інтенсивності звука); спектрографію (для

фіксування загальної акустичної картини звука).

У прикладній фонетиці розрізняють первинні та вторинні звукові сигнали. Первинні – безпосередньо породжені мовцем, а вторинні – ті ж первинні, але додатково пропущені через канали електроакустичних пристроїв (наприклад, мікрофони, гучномовці, підсилювачі тощо).

Одні й ті самі ж мови можуть мати варіанти – алофони (наприклад, у наголошеному й ненаголошеному складах, в літературному й діалектному мовленні). Крім того, кожен мовець генерує лінгвістичні звуки з індивідуальними відмінностями. Та навіть звуки одного мовця, вимовлені в різні моменти часу, різняться тривалістю, силою тощо.

У комп'ютерній лінгвістиці на перше місце за актуальністю все більше виступає акустичний аспект дослідження звуків мови.

Мовні звуки, як і будь-які інші, характеризуються фізичними характеристиками – частотою, інтенсивністю, тривалістю, а також спектральним складом (у мовознавстві спектральний склад ще називають формантним).

Частота звука

Найпростішим типом звуку є звук, в якому тиск у кожній точці простору змінюється за синусоїдним законом, тобто здійснює гармонічні коливання з певною частотою. Частота – це кількість коливань певної точки звукової хвилі в секунду. Одному циклу коливання в секунду відповідає величина 1 Гц (1/с).

Однією з характеристик звука є частота, з якою згущені та розріджені ділянки повітря діють на органи слуху. Частоту звука вимірюють у кількості коливальних періодів за 1 с. Одиницею вимірювання є герц (Гц).

Людина чує звук з частотами від 16 Гц до 20 кГц. Деякі тварини можуть чути звуки з частотою, нижчою від 16 Гц, інші – з частотою понад 20 кГц.

Діапазон від 16 Гц до 20 кГц називають чутним діапазоном. Звуки з частотами до 16 Гц називаються інфразвуком, понад 20000 Гц – ультразвуком. Звуки з частотою 109−1013 Гц називають гіперзвуком. Людське вухо сприймає та розрізняє частоту звукових коливань як висоту звуку або тон. Частоти, вищі за 20 КГц, органи слуху людини не сприймають. Так само органи слуху людини не сприймають тони коливання, нижчі за 20 Гц (людина сприймає їх як рипіння, биття, стукання). Представники тваринного світу здатні

сприймати звуки в значно ширшому діапазоні частот. Кажани для орієнтації та знаходження здобичі використовують звуки в діапазоні 20-120 кГц. Дельфіни, наприклад, використовують для орієнтації та полювання звуки з частотою понад 100 кГц. Значно вищі 20 кГц частоти здатні сприймати і собаки.

Інтенсивність звука – густина потоку звукової енергії. Найменша інтенсивність звуку, яку ще може сприймати вухо людини (поріг чутності), становить 10^{-16} Вт/см². Інтенсивність звука вимірюють силою тиску, що діє на одиницю площі.

Цікаво, що суб'єктивне сприйняття людиною збільшення інтенсивності звука насправді може бути викликане зростанням його висоти.

Тривалість звука у фонетиці вимірюють у мілісекундах, тобто одній тисячній секунди. Наприклад, тривалість пауз у мовному потоці більше ніж 100 мс (0,1 с) дає людині змогу відокремлювати звуки, а менша – сприймати їх як один.

Спектральний склад звука – це сукупність простих гармонійних хвиль, на які можна розкласти звукову хвилю. Спектральний склад звука виражає його частотний (спектральний) склад і виходить в результаті аналізу звуку.

Екпериментально встановлено, що мовні звуки української мови, наприклад, містять: [а] – одночасно приблизно 650,1250 і 2200 Гц; [і] – одночасно приблизно 250, 2750 і 3400 Гц.

Отже, тим лінгвістам-прикладникам, які планують займатися прикладною фонетикою, наприклад, створенням систем аналізу чи синтезу усного мовлення, вкрай потрібні знання з акустики, яка є розділом фізики.

Раніше для аналізу звуків використовували кімографи, в яких звук фіксувався самописцем на смужці паперу, що прокручувалася. У наш час для візуального зображення звука використовують звукові редактори – спеціальні програми, що працюють на комп'ютерах.

Пристроями, які дають змогу вводити в комп'ютери звук та виводити його, є звукові плати, вмонтовані в системні блоки комп'ютерів. На таких звукових платах є спеціальні мікросхеми, що перетворюють коливання повітря, отримані через мікрофон, у цифрові коди й, навпаки, цифрові коди – в коливання мембрани гучномовців, під'єднаних до звукових плат. Деякі з таких мікросхем іноді монтують навіть безпосередньо на материнських

платах комп’ютерів.

У звукових редакторах звук може відтворюватися як у двовимірному (відтворюється тривалість звука та його амплітуда), так і в тривимірному просторі (відтворюється тривалість звука, його частота й інтенсивність).

Серед звукових редакторів можна виділити: а) звукові редактори загального призначення (наприклад, для опрацювання музики) та лінгвістичного (для опрацювання саме усного мовлення); б) звукові редактори з відображенням і без відображення спектра звука.

Для досліджень артикуляції органів мови використовують також рентгенокінографування – фіксування на плівці рухів органів мовлення людини. Під час таких експериментів рентгенокінографування проводять з потрібною кількістю кадрів на секунду, а далі, в разі потреби, сповільнено повторюють процес утворення звука й створюють, наприклад, моделі його утворення.

Перед сучасною прикладною фонетикою стоять такі основні завдання: створення комп’ютерних баз даних еталонних звуків мови (з метою їх майбутнього використання для розпізнавання усних текстів у комп’ютерних лінгвістичних системах); створення спеціалізованих звукових лінгвістичних редакторів; створення спрощених систем автоматичного аналізу усного мовлення для дистанційного керування робототехнічними пристроями; створення систем автоматичного аналізу усного мовлення для його перетворення в писемний текст, тобто створення електронних секретарів; створення систем автоматичного синтезування усного мовлення (перетворення писемного тексту в усний, тобто створення електронних секретарів-декламаторів).

Поняття про письмо

Писемна мова є похідною (вторинною) стосовно усної. Письмо – це графічна знакова система, призначена для фіксації усного тексту на площинному носії інформації. Графічна знакова система – це сукупність (множина) графем, а графема – мінімальний двомірний графічний символ (рисунок, піктограма), що позначає одиницю мови. Найчастіше графеми позначають: літера – фонему; літера чи лігатура (кілька літер, об’єднаних в одну) – частину або цілий склад; ієрогліф – частину слова або ціле слово. Прикладом графічних знакових систем є множини літер (абетки), множини лігатур чи множини ієрогліфів.

Письмо дає змогу передавати тексти як у часі, закріплюючи їх на матеріальних носіях інформації, так і на потрібну віддаль у просторі.

Розділ прикладного мовознавства, який досліджує письмо, називають графікою.

На початковій стадії свого розвитку «письмо» було не знаковим, а предметним. Типовим прикладом такого письма є вузликове. У давнину вузликове письмо було в широкому вжитку в китайців, мексиканців, персів та інших народів. Здатність предметів символізувати поняття була різнобічно використана стародавніми племенами. Це представлено в таких видах індійського письма як вампум і кіпу.

Вампум – це раковинні буси ірокезьких племен; вони слугували не тільки прикрасами, адже їх використовували і як предметні «записи». Особливо важливою була символіка кольору: білий колір означав мир, здоров'я, благополуччя; чорні буси сповіщали про смерть вождя; червоні – про війну.

Кіпу – це вузликове письмо інків. Кожний кіпу складався з основного шнура, до якого прикріплювалися вузли та сплетіння у вигляді бахроми: вузли та сплетіння різного кольору та розташування були замітками про різні адміністративні та господарські розпорядження.

Спеціально виготовлені шнури та буси виступали зоровими засобами спілкування. Ще більш умовний характер отримували зарубки, татуювання, родові та племінні знаки, знаки власності. Вони наносилися на предмети побуту, на стіни печер, скелі тощо. Цей печерний та наскальний живопис і був одним з найдавніших видів письма.

Піктографія. Предметна символіка була передісторією письма. Першим історичним типом письма була піктографія, тобто малюнкове письмо. Піктограми – одиниці такого письма – видряпувались, а потім і малювались на стінах печер, на скелях і каменях, на рогах і кістках тварин, на бересті.

Класичним прикладом піктографічного письма може слугувати прохання семи індійських племен, подане у січні 1849 р. президенту США. Племена представлені їх тотемами – племінними назвами – журавель, три куниці, ведмідь, морська людина і морський кіт. Спереду – вождь журавлиного племені. Невелика рисочка від очей глави делегації направлена на президента, а лінія,

яка іде від його очей назад до озер, вказує на предмет його прохання. Племена просять дозволити їм переселитися з верхнього озера (воно зображено у вигляді довгої блакитної лінії, над якою розміщені всі фігури) на нижні озера, які зображені знизу. Щоб показати, що всі члени племені погоджуються зі своїм вождем у цьому проханні, – очі та серця всіх шести фігур з'єднані з очима та серцем журавля. Піктографічне письмо ще зберігає зв'язок з живописом, виражаючи думку через символічне осмислення зображених облич, предметів, явищ. Зараз піктографія не використовується як основний вид письма. Але піктограми широко використовуються на вітринах, рекламах, як дорожні знаки тощо.

Ідеографія. Утративши зоровий образ та перетворившись у немотивований знак, піктограма перетворилася на ідеограму, а піктографія стала ідеографією. Вона виникла тоді, коли виникли країни, для яких письмо було необхідним для управління та обліку державного та храмового господарства, для запису законів та культових текстів. Найдавнішими ідеографічними системами були єгипетська, шумерська, китайська, ацтекська, майя.

Прикладом найдавнішого єгипетського письма може бути шиферна табличка фараона Нармера (біля 3 тис. років до н.е.). Фараон зображений у короні Верхнього Єгипту, він заніс булаву над головою переможеного ворога. Цей малюнок містить піктографо-ідеографічний надпис: шість стеблин папірусу; над ними сокіл, який тримає у лапі мотузку, яка пронизана через губи людської голови; прямокутник з хвилястими лініями; зображення риби та булави. Елементи даної композиції не піддаються буквальному осмисленню. Необхідна дешифровка, тобто прочитання піктограм, ідеограм та ключів, які вказують на слова (у даному випадку ім'я фараона). Радянський єгиптолог Б. А. Тураєв дешифрував надпис таким чином: сокіл – символ фараона: він тримає мотузку, пронизану через губи людської голови, і це означає, що фараон вивів як полонених людей завойованої ним країни. Полонених було 6000, тому що стеблина папірусу означала цифру 1000. Хвилясті лінії у прямокутнику як символ води означають, що завойована країна розташована на березі моря. Ім'я фараона, передано зображенням риби і булави, читається «нармер». Весь напис можна прочитати так: «Фараон Нармер вивів із завойованої ним морської країни 6000 полонених». Щоправда, інші єгиптологи, наприклад, А. Гардинер, сумніваються у

фонетичному читанні зображеної риби та булави, а стеблини папірусу розуміють як символ Нижнього Єгипту. Тому надпис вони читають так: «Цар веде полонених із Нижнього Єгипту».

Аналіз найдавнішого ідеографічного письма показує, що ідеограма виникла на базі малюнка і піктограми, особливо при зображенні конкретних понять, – таких, наприклад, як «будинок», «людина», «вода».

Найбільш стійким та розповсюдженим до цього часу ідеографічним письмом є китайське, чому сприяла особлива історія Китаю і односкладний характер китайських слів

Ідеографічний принцип письма отримав широке застосування при записах рахунку та алгебраїчних понять, що й привело до створення цифр та наукової символіки (алгебраїчної, хімічної, логічної тощо). Слово цифра арабського походження: цим словом араби називали нуль (буквально: пусте місце). Найдавніші цифри з'явилися у Вавілонії та Єгипті. При виникненні алфавітів цифри позначалися буквами; наприклад, буква альфа в давнину позначала цифру 1. Зараз найбільш розповсюдженими цифровими системами є арабська та римська. Арабські цифри виникли в Індії; у Європу вони були занесені арабами (звідси їх вторинна назва).

Силабічне (складове) письмо. Наступним кроком у вдосконаленні писемної мови був перехід від позначення слів до позначення складів. Таке письмо вже повністю передавало усну вимову слова. При цьому виділяли такі типи складів: приголосний + голосний; голосний + приголосний; приголосний + голосний + приголосний; приголосний + приголосний + голосний; приголосний + голосний + приголосний + голосний (два склади одночасно).

Іноді в таких письменах виділяли також окремі приголосні чи голосні. Кількість знаків у таких письменах досягала 100–300.

Прикладом такого письма є деванагарі (санскрит). Зараз таке письмо використовують в Індії.

Літерно-звукове письмо. Наступним кроком у розвитку письма був перехід від позначення складів до позначення звуків. Це реалізувало принцип: позначати одним знаком приблизно одну фонему. Історично відомі два різновиди цього письма – консонантне й вокалізоване.

У консонантному письмі символами позначали переважно приголосні звуки. Прикладами служать фінікійське,

давньоєврейське, арамейське, арабське й древньогрецьке письмо. Час виникнення цього письма – 1,5 тис.р. до н. е. Першими таке письмо (22 приголосні звуки) застосували фінікійці. При такому письмі голосні звуки доводилося вгадувати. Потім для полегшення читання над чи під приголосними почали ставити спеціальні знаки (наприклад, як в арабському письмі – одну, дві чи три крапки).

У вокалізованому письмі символами позначали як приголосні, так і голосні звуки. Вперше у VIII ст. до н. е. на основі фінікійського таке письмо застосували в Греції. Далі давньогрецьке письмо розділилося на західне (у VII ст. до н. е. виникли італійська й латинська абетки) й східне (у V ст. до н. е. – класична грецька абетка, а далі – візантійське письмо).У часи середньовіччя абетки поширювалися разом з релігіями (наприклад, арабська абетка – в країнах ісламу; латинська – в країнах християнства).

В Європі латинська абетка набула поширення разом із поширенням католицизму. При цьому виникали певні труднощі: для позначення відсутніх у латинській мові звуків доводилося використовувати бі-, три- й тетраграми, а також літери з діакритичними знаками.

У IV ст. н. е. були створені абетки для мов народів Кавказу.

Цікаво, що до появи кириличного письма у слов'ян існувало інше письмо – глаголиця. Вона повністю збігається з кирилицею за кількістю літер, їх порядком в абетці й відповідністю фонемам. За однією з гіпотез, глаголиця була створена Кирилом (Константином-філософом) ще до поїздки Кирила й Мефодія в Моравію. Глаголицю широко використовували в середині IX ст. Іноді її використовували в Стародавній Русі (є письмові пам'ятки X ст.). У Хорватії в церковних книгах глаголицю використовують дотепер.

Види систем письма

Існують два види письма: традиційне й нетрадиційне. *Традиційне письмо* – це письмо, призначене для використання потенційно всіма членами суспільства. Звернемо увагу на те, що деякі мови мають по дві чи навіть по три паралельні системи письма. Так, сербська мова паралельно використовує як кирилицю, так і латиницю; японська мова одночасно використовує ієрогліфічне письмо, літерне письмо для слів власного походження (хірагана), а також літерне письмо для слів іноземного походження (катакана).

Спеціальне письмо призначене для використання лише якоюсь

частиною суспільства чи з якоюсь певною метою. Такими частинами суспільства можуть виступати, наприклад, незрячі чи люди з вадами органів мовлення. Цілі, для яких можуть використовувати спеціальні системи письма, є такі: зафіксувати вимову усного тексту на носії інформації; передати писемний текст за допомогою іншої абетки; обмежити коло тих, хто може читати писемний текст; зробити абетку зручною для використання в технічних каналах передавання інформації тощо.

Останнім часом у системах письма з'являються суттєві доповнення. Так, використання комп'ютерів сприяло тому, що графеми мають зафіксовані міжнародними стандартами цифрові коди (раніше цифрові коди графем використовували тільки в спеціальних системах письма), а носії інформації, на яких записують тексти, можуть бути й тривимірними.

З часом письмо в різних мовах удосконалювалося, зокрема доповнювалася графіка, встановлювалися орфографія та напрям письма. Саме ці три компоненти – графіка, орфографія, напрям письма – становлять систему письма.

Графіка – це сукупність усіх символів, які використовують у певний часу писемному мовленні якоїсь мови. Наприклад, до складу графіки української мови, крім абетки, входять пробіл, розділові знаки, арабські й римські цифри (вони є ієрогліфами!), апостроф, знак наголосу, а також інші допоміжні знаки, яких дуже багато, особливо в науковій літературі.

Основною складовою графіки є абетка. *Абетка* – це упорядкована множина графем, які позначають звуки певної мови й призначені для графічного відображення звукових образів слів. У принципі, зовнішній вигляд літер є довільним, проте з погляду можливості автоматичного розпізнавання вкрай бажана їх суттєва відмінність (наприклад, в українській абетці важко розрізняти ш і щ, г і ґ).

Абетка задає строго визначену відповідність між кожною літерою та позначуваним нею звуком. Крім того: один звук може позначатися й кількома літерами (наприклад: **[п]** може позначатися й літерою б, й літерою п: *дуб, порошок);* одна літера може позначати різні звуки (наприклад, у слові *яблуко* – два звуки **[ja],** а в слові *няня* – один звук **[а]);** кілька звуків можуть позначатися однією літерою; кілька літер можуть позначати один звук.

У низці мов для позначення потрібних звуків через відсутність

необхідної кількості літер (наприклад, у латинській абетці) до наявних додають діакритичні знаки чи інші графічні елементи. Такі методи свого часу були широко використані в західнослов'янських мовах – чеській, польській тощо.

У словах літери записують у тому порядку, в якому йдуть у слові відповідні їм звуки.

Ієрогліфічні мови замість абеток мають упорядковані списки ієрогліфів. Так, графіка китайської мови (близько 24 тис. основних і 34 тис. рідковживаних) базується на списку ієрогліфів, що має 214 ключів, які вказують розташування кожного з ієрогліфів у словнику.

Орфографією називають історично усталену й загальноприйняту систему правил фіксування усних текстів певної мови на письмі. В основу орфографії кладуть кілька принципів.

1. Фонетичний – кожен звук позначають відповідною літерою, а далі діють за правилом: як вимовляють, так і пишуть.
2. Морфологічний – морфеми завжди пишуть однаково (наприклад: *дуб – дуби).* Проте можливе порушення однотипного написання одних і тих же.
3. Традиційний – написання приймають у такій формі, як воно утвердилося історично.
4. Диференційний. Згідно з цим принципом для різних позначуваних об'єктів чи значень використовують різні написання, хоча слова мають однакове походження.
5. Транслітераційний *(Берлін)* і транскрипційний *(Нью-Йорк)* – для чужомовних слів.

З часом правопис зазнає змін. Так, зараз в Україні використовують четвертий правопис. До впровадження планують п'ятий, що повинен передбачати 10 змін. Плануються зміни й до німецького правопису.

Системи письма різняться також *напрямом.* Відомі такі напрями письма: справа наліво (арабське); зліва направо й справа наліво (зі зміною для кожного слова, записаного одне під другим у стовпчик) бустрофедон (імовірно, давньогрецьке); зліва направо (латинське, кириличне); згори вниз (китайське).

Для оцінювання якості графіки використовують різні критерії. Так, співвідношення кількості літер і кількості звуків мови визначає оптимальність алфавіту (А). Її підраховують за формулою: $A = L / S$, де L – кількість літер в алфавіті певної мови, а S – кількість звуків у

цій мові.

Чим ближче А до 1,0, тим оптимальнішим є алфавіт. З часом розбіжність між кількістю літер і кількістю звуків, як, наприклад, в англійській мові, може зростати.

Коли внаслідок того, що різні літери позначають один і той самий звук, можливі різні варіанти написання слова, тоді, як вказувалося, використовують норми орфографії. Враховуючи наявність таких фактів, лінгвісти встановили такий показник як надлишковість графіки (Н). Цей показник визначають за формулою: **H = G / S** де G – кількість графем у певній мові, а S – кількість звуків у ній.

Чим більша за 1,0 величина Н, тим більша надлишковість графіки.

Здавна люди використовували для письма багато різних пристроїв. Основними серед них можна назвати: стило (невеличка паличка з загостреним кінцем) і глиняні чи з іншого матеріалу таблички; перо (ручка) чи олівець і папірус, пергамент або папір; клавіатура й папір, фотопапір, фотоплівка, друкарська форма.

З другої половини XX ст. для традиційного письма дедалі частіше використовують електронні пристрої постійного і тимчасового запам’ятовування.

Першими засобами для зображення символів були рука, палиця і камінь. Рукою і палицею можна було малювати на піску, а каменем – на скелі.

Близько 4000 років до н.е. людина почала використовувати змочені глиняні дощечки. При цьому ручкою служила дерев’яна або бронзова паличка, або кістка.

Близько 3000 до н.е. єгиптяни винайшли форму письма у вигляді ієрогліфів.

Близько 1300 до н.е. римляни стали використовувати письмо по воску. Віск заливався в дерев’яні таблети. У цей час дано назву інструменту – стилус (stylus). Стилус виготовлявся з металу. Коли запис ставала не потрібна, вона стиралася за допомогою плоского зворотного кінця стилуса.

У цей час в Китаї використовували стилуси, виготовлені з бронзи.

Письмо з воску практично без змін проіснувало 18 століть поки англосакси не винайшли пергамент. З появою пергаменту, що використовується для виготовлення рукописних книг, люди

продовжували вживати воскові дощечки для щоденних записів та макетування книг.

Європейці (в першу чергу іспанці) виявили, що при використанні певним чином заточеного гусячого пера для письма по пергаменту, можна змінити і стиль письма – зробити його прописним і похилим. Так 14 століть тому були придумані прописні букви. До цього при листі використовували лише великими літерами.

Гусяче пір'я проіснували рекордно тривалий час – до кінця 18 століття.

Приблизно в 1790 австралійці і французи винайшли грифель для олівця. Тоді й виникла індустрія інструментів, що пишуть.

У 1803 році було запатентовано металеве перо для ручки. Однак лише через 27 років сталеві кінчики для пір'я отримали визнання на ринку. До кінця 19-го століття ручки з металевим пером повністю витіснили недовговічні, що вимагають частої заміни гусячі пір'я. Усі спроби винахідників зробити довгу друкарську ручку закінчувалися невдачею.

У 1884 році страховий агент Левіс Едсон Ватерман увійшов в історію, як винахідник ручки, що заправляється чорнилом (спочатку збоку, за допомогою спеціальної піпетки). Йому так багато доводилося писати на роботі, що це підштовхнуло його до винаходу.

Транскрибування – це вид спеціального письма, за допомогою якого фіксують вимову усних текстів. Запис можуть здійснювати з різним ступенем точності, що залежить від потреб того, хто записує. Транскрибування розрізняють залежно від одиниць мови, які записують: для опису звуків використовують звукове, а для опису синтагм і фраз – інтонаційне транскрибування.

Транскрибування виникло з потреби усного спілкування тих, хто вивчає мову, з її носіями. Адже писемна мова забезпечити такої можливості не може, оскільки не фіксує вимову одиниць мови.

Звукове транскрибування. У звуковому транскрибуванні виділяють фонетичне, фонематичне й практичне. За допомогою фонетичного транскрибування фіксують усі основні особливості мовлення (наприклад, слово української мови *кожух* записують як [ко^{у}ж'ух]). За допомогою фонематичного транскрибування вказують лише зафіксовані в мові фонеми (наприклад, для української мови це тільки б голосних і 32 приголосних звуки), а не

реально вимовлені в текстах. Тому, наприклад, те саме слово *кожух* запишуть лише як <кож'ух>. Практичне транскрибування застосовують для відтворення особливостей вимови чужомовних слів літерами своєї абетки (наприклад, англійське слово *mother* записують як *мазе).* Воно має дуже обмежену сферу вживання (наприклад, туристичні путівники, газетні публікації), оскільки не відповідає критерієві точності передання звуків мови, яку транскрибують.

При потребі звукова транскрипція може передавати ще й висоти тонів, що, наприклад, важливо для китайської мови.

Для фонетичної транскрипції використовують позначення Міжнародної фонетичної асоціації (МФА), виконані на базі латинської абетки.

Транскрипція, за допомогою якої вказують висоти звуків (іноді її називають «мелодійною»), уніфікованої системи позначень не має.

Інтонаційне транскрибування. При інтонаційному транскрибуванні синтагми розділяють однією (/), фрази – двома (//), а кінець тексту позначають трьома (///) косими лініями.

Застосування транскрибування. Фонетичне транскрибування використовують в орфоепічних словниках. Такі словники служать джерелом нормативної літературної вимови. Наприклад, в українській мові нормативною є тверда вимова шиплячих [шчо], [чого] (пом’якшення шиплячих не допускається).

Практичне транскрибування використовують для найпростіших типів перекладних словників-розмовників, путівників, перекладів на рідну мову іноземних власних назв, термінів, номенклатур, топонімів тощо. При цьому допускається незвичне поєднання літер. В Україні таблиця для практичного транскрибування (передавання української ономастики – найрізноманітніших власних назв – засобами англійської мови) на основі національної абетки затверджена Міністерством юстиції України 1996 р.

Інтонаційне транскрибування широко застосовують у ЗМІ для розмічування текстів дикторам радіо й телебачення, а також для акторів театру й кіно.

Транслітерування – це вид спеціального письма, за допомогою якого тексти, написані однією графікою, політерно відтворюють за допомогою іншої графіки. У найпростішому випадку графікою

виступає абетка. Транслітерування допускає умовне використання літер, доповнення абетки додатковими літерами, а також використання діакритичних знаків. Транслітерування за допомогою латинської абетки називають романізацією (наприклад, існує спеціальний спосіб запису японських текстів літерами латинської абетки – ромаджі).

Транслітерування виникло в XIX ст. в Пруссії через потребу укласти бібліографічні описи видань іноземними мовами (слов'янських, країн Близького й Далекого Сходу, Індії та ін.).

Правила транслітерування розробляє Міжнародна організація стандартизації (МОС). Іноді стандарти МОС і конкретної держави можуть не збігатися, в країні одночасно можуть використовувати кілька різних систем транслітерування тощо. Наприклад, для російської мови відомо близько 20 таких систем, хоча офіційно затверджена лише одна.

Стандарту на транслітерування літер української абетки латинською немає.

Стенографія – це вид спеціального письма, за допомогою якого усний текст можна записувати в кілька разів швидше, ніж за допомогою звичайного письма.

Зростання швидкості (від чотирьох до семи разів) забезпечується за рахунок спрощення написання літер (знаряддя писання – ручка, олівець тощо – повинно переміщатися тільки вперед), а також за допомогою інших прийомів. У стенографічному письмі використовують елементи символів традиційних абеток: крапки, лінії (прямі й нахилені праворуч), овали, частини овалів, сполучення всіх цих елементів.

Стенографія зародилася в Єгипті, отримала розквіт в античних Греції та Римі. Автором давньоримської стенографії (I ст. до н. е.) вважають Тирона – раба й секретаря відомого оратора Ціцерона. Саме слово *стенографія* з'явилося 1602 р. (цим словом в Англії назвав свою працю «Мистецтво скоропису, або стенографія» Джон Уілліс). Стенографія в перекладі зі старогрецької означає «вузькопис, тіснопис». У 1933 р. в Лондоні світова громадськість відзначила 2000-ліття стенографії. У першій половині XX ст. існували спеціальні стенографічні друкарські машинки.

У СРСР 1933 р. на знаки стенографування для російської мови було введено стандарт. У 1967–1968 рр. він був удосконалений, внаслідок чого в ньому використовувалося 25 знаків: 20 – для

позначення приголосних і 5 – для позначення голосних. Українська стенографія виникла лише в 20-30-х роках ХХ ст.

У наш час, завдяки наявності диктофонів (не тільки стрічкових – аналогових, а й цифрових), потреба в стенографуванні звужується, однак його використовують, наприклад, у дипломатії, під час синхронного перекладу тощо.

Це система спеціальних значків: геометричних (використовуваних у різних системах стенографії в Англії й Франції) та елементів літер рукописного письма (застосовуваних головно в Німеччині й по слов'янських країнах), пристосована для швидкого запису за складами, словотвірними морфемами й словами усної мови.

Системи письма для незрячих – це вид спеціального письма, який дає змогу позначати літери абетки знаками, які людина може сприймати тактильними аналізаторами (наприклад, пальцями).

У XVII–XIX ст. для незрячих було розроблено десятки різних систем письма, які базувалися на одному принципі – бути одночасно зручними як для незрячих, так і для зрячих людей. На противагу цьому француз Луї Брайль (1809-1852 рр.) запропонував систему письма для незрячих, зручну саме для них. Хоча спочатку система письма Луї Брайля не знайшла прихильників, з кінця XIX ст. вона стала загальноприйнятою (міжнародною).

Луї Брайль запропонував не тільки шрифт, а й приладдя для нього – сторінку у вигляді таблиці (така сторінка містить 18 рядків по 24 клітинки в кожному, тобто 432 знаки на сторінці); спеціальні рамки з дротиків, які дають змогу писати рівно, по рядках; спеціальний «олівець»-шило – ним роблять у цупкому папері дірки, причому справа наліво, тоді перегортають сторінку й читають літери пальцями зліва направо.

Крім системи письма, Луї Брайль розробив для незрячих також систему нотопису.

Для кодування літер, цифр та інших знаків брайлівська система письма використовує шість опуклих точок (два стовпці по три точки), що дає змогу кодувати 64 знаки.

У шрифті Брайля існують спеціальні позначення для цифр (арабських та римських), знаків пунктуації, грецьких літер, математичних знаків, функцій, хімічних знаків тощо. Частину з них утворюють за допомогою сполучення кількох знаків (двох, трьох, чотирьох).

Мова жестів – спеціальне письмо, яке дає змогу позначати літери, а також цілі слова жестами. Мову жестів для обміну інформацією використовують як люди з вадами голосових зв'язок і слуху, так і люди без таких вад. Відповідно розрізняють: а) мову жестів для людей без вад органів мовлення (наприклад, в австралійських племенах вдови після смерті чоловіків рік розмовляють лише мовою жестів); б) мову жестів для людей з вадами органів мовлення – глухих чи глухонімих (таких людей від 0,4 до 1,5%). Мова жестів за своїми можливостями не поступається звуковій мові, хоча соціально має нижчий статус.

Структурно мова жестів складається з двох компонентів: множини тисячі-півтори слів,для позначення яких використовують окремі жести (як правило, це найчастіше вживані слова), а також решти слів, для яких використовують пальцеву абетку (в цьому випадку пальцями позначають окремі літери). Розрізняють одноручну та дворучну абетки.

Цікаво, що мову жестів не залежать від конкретних звукових мов. Так, мова *амслен* є спільною для США, Канади й Франції, де використовують дві різні мови (англійську та французьку), але не Англії (там мова жестів зовсім інша, хоча традиційною є англійська).

Одиницею мови жестів є жест, або херема (від грец. рука*).* Є три класи херем: ті, що вказують на місце виконання жесту; ті, що вказують на конфігурацію руки; ті, що вказують на характер руху. Кількість херем приблизно збігається з кількістю фонем. Наприклад, в амслені – 55, у шведській мові жестів – 64, у південнофранцузькій – 53 хереми. Для передання граматичної інформації такі мови використовують тримірність простору.

Перелічимо деякі відмінності між мовою жестів і звуковою мовою. Якщо в звуковій мові лексичну й граматичну інформацію подають зі зсувом у часі (спершу префікс, далі корінь, потім афікси й флексія), то в мові жестів їх подають паралельно. У мові жестів існує свій поділ слів за частинами мови. Так, прикметники належать до одного з різновидів предикатів, тобто дієслів; для дієслів вказується кількість об'єктів. Описуваних мовців автор тексту поміщає в різні точки простору. Крім жестів, використовують також немануальні компоненти (міміку), наприклад для розрізнення омонімів є мімічні прислівники.

У наш час створено комп'ютерні системи, які дають змогу

перетворювати мову жестів в усний текст. Для використання таких систем потрібно одягнути спеціальні рукавички з детекторами, під'єднаними до комп'ютера.

Криптографування (грец. *таємне писання)* – це вид спеціального письма, який дає змогу записати зафіксований природною писемною мовою текст в такий спосіб, який максимально ускладнює його політерне розпізнавання.

Криптографування складається з двох протилежних процесів – шифрування й дешифрування. Шифрування – це зміна в тексті зображення символів, їх кодів чи порядку за наперед заданими правилами. Такі правила називають ключем. Використовують криптографування тоді, коли хочуть зробити текст доступним (зрозумілим) лише для обмеженого кола отримувачів. Дешифрування – це відновлення в тексті за допомогою ключа первинного зображення символів, їх кодів чи порядку. Дешифрування можливе й без знання ключа, проте в такому випадку воно вимагає набагато більших витрат часу на відновлення тексту.

Розрізняють дешифрування текстів, які були спеціально зашифровані (наприклад,у дипломатії, розвідці), й текстів, які не були спеціально зашифровані (наприклад, написані мертвими мовами). У коло дослідження лінгвістики входить саме останнє дешифрування – так зване лінгвістичне. У лінгвістичному дешифруванні розрізняють варіанти, коли: відомі графіка й мова, але невідомий зміст тексту, тобто значення слів; невідома графіка тексту, але відома його мова; відома графіка тексту, але невідома його мова; невідомі ні графіка, ні мова тексту.

Під час лінгвістичного дешифрування розпізнати текст допомагає наявність паралельних перекладів, знання власних імен та історичних подій. У результаті лінгвістичного дешифрування створюється ключ, який має містити перекладний словник і граматику мови.

Крім лінгвістичного, виділяють також дешифрування загальне (наприклад, у біології – дешифрування генів) і технічне (наприклад, старих карт, нотних записів, військових текстів). За іншою класифікацією, виділяють дешифрування прикладне (знаходження одного ключа для конкретного тексту) й загальне (розроблення загальних способів знаходження ключів).

Історія криптографії налічує близько 4 тисяч років. Як

основний критерій періодизації криптографії можна взяти технологічні характеристики використовуваних методів шифрування.

До нашого часу криптографія займалася виключно забезпеченням конфіденційності повідомлень (тобто шифруванням) – перетворенням повідомлень із зрозумілої форми в незрозумілу і зворотне відновлення на стороні одержувача, роблячи його неможливим для прочитання для того, хто перехопив або підслухав без секретного знання (а саме ключа, необхідного для дешифровки повідомлення). В останні десятиліття 21 ст. сфера застосування криптографії розширилася і включає не лише таємну передачу повідомлень, але і методи перевірки цілісності повідомлень, цифрові підписи, інтерактивні підтвердження, та технології безпечного спілкування тощо.

Найперші форми тайнопису вимагали не більше ніж аналогу олівця та паперу, оскільки в ті часи більшість людей не могли читати. Поширення писемності, або писемності серед ворогів, викликало потребу саме в криптографії. Основними типами класичних шифрів є перестановочні шифри, які змінюють порядок літер в повідомленні, та підстановочні шифри, які систематично замінюють літери або групи літер іншими літерами або групами літер. Прості варіанти обох типів пропонували слабкий захист від досвідчених супротивників. Одним із ранніх підстановочних шифрів був шифр Цезаря, в якому кожна літера в повідомленні замінювалась літерою через декілька позицій із абетки. Цей шифр отримав ім’я Юлія Цезаря, який його використовував, зі зсувом в три позиції, для спілкування з генералами під час військових кампаній.

Поява цифрових комп’ютерів та електроніки після Другої світової війни зробило можливим появу складніших шифрів. Більше того, комп’ютери дозволяли шифрувати будь-які дані, які можна представити в комп’ютері у двійковому виді, на відміну від класичних шифрів, які розроблялись для шифрування письмових текстів. Це зробило непридатними для застосування лінгвістичні підходи в криптоаналізі. Багато комп’ютерних шифрів можна характеризувати за їхньою роботою з послідовностями бінарних бітів (інколи в блоках або групах), на відміну від класичних та механічних схем, які, зазвичай, працюють безпосередньо з літерами. Однак, комп’ютери також знайшли застосування у

криптоаналізі, що, в певній мірі, компенсувало підвищення складності шифрів.

Азбука Морзе

Ще один спосіб кодування традиційного письма, який донедавна використовували в електричних каналах зв'язку, називають за прізвищем її автора азбукою Морзе.

Передавати інформацію каналами електричного зв'язку в Європі почали в першій половині XX ст. Найдосконаліший спосіб кодування знаків (на основі їх частотності), а також найдосконаліший апарат для їх передавання 1838р. запропонував англієць Самюель Морзе (1791–1872 рр.).

Ефективність передавання текстів за допомогою таких кодів у другій половині XIX і першій половині XX ст. досліджували відомі математики – Вільям Томсон (лорд Кельвін), Томас Едісон, Олександр Белл, Анрі Пуанкаре. Наприкінці 40-х років проблемою передавання інформації каналами телеграфу особливо продуктивно займався К. Шеннон, який, базуючись на принципах роботи телеграфу, заклав основи теорії інформації – складової кібернетики.

Знаки в азбуці Морзе кодують короткими («крапками») й довгими («тире») електричними сигналами. При цьому крапку вважають одиницею тривалості. Тривалість тире повинна дорівнювати тривалості трьох крапок. Пауза між сигналами в літері – одній крапці, а пауза між літерами – трьом крапкам.

В азбуці Морзе існують також позначення для розділових знаків (крапка, кома, знак оклику, знак питання, крапка з комою, двокрапка, дужки, дефіс, лапки) та спеціальних сигналів (границя розділу, стирання попереднього знака, початок передавання, готовність до прийому, початок дії, закінчення передавання тощо).

З поширенням комп'ютерної техніки (відповідно, цифрових кодів) використання азбуки Морзе (аналогових сигналів) скорочується. Зараз нею найчастіше користуються радіоаматори.

З виникненням комп'ютерів для організації їх спілкування з людьми постало завдання – створити спеціальні шрифти (шрифтокомплекти), в яких літери позначалися б не графічними (як у традиційному письмі), а цифровими образами (цифровими кодами). До кожного такого цифрового коду «причіпляли» відповідне графічне зображення знака.

У створенні шрифтокомплектів можна виділити три етапи.

На першому етапі кожен знак (у тому числі кожну літеру)

кодували вісьмома бітами, що відповідає одному байту. Оскільки байт може позначати не більше 256 різних комбінацій двійкових чисел, то в такий шрифтокомплект могло входити не більше 256 знаків. Літери мали лише одне накреслення (як на друкарській машинці). Такі шрифтокомплекти записували в спеціальні мікросхеми дисплеїв і друкарок назавжди.

На другому етапі, як і на першому, кожен знак продовжували кодувати одним байтом, проте графічне зображення знаків уже могло урізноманітнюватися (як у шрифтах Courier New, Times New Roman, Ariel). Такі шрифти вже не записували назавжди в спеціальні мікросхеми дисплея чи друкарки, а використовували як звичайне інформаційне забезпечення комп'ютера, котре, як і будь-які програми, можна встановити, а в разі потреби – видалити. Для подання в шрифтокомплектах інформації про графічне накреслення знака було розроблено спеціальні стандарти. Таких оцифрованих шрифтів було розроблено дуже багато.

На третьому етапі, що розпочався з появою Інтернету, для забезпечення спілкування людей, що обмінюються інформацією на різних мовах (наприклад, на базі латинської абетки та ієрогліфів одночасно), постала потреба суттєво розширити кількість знаків у шрифтокомплектах (переважно літер, лігатур та ієрогліфів). Наслідком пошуків стали шрифтокомплекти, в яких один знак кодували не одним, а двома байтами (два байти дають змогу кодувати більше 60 тис. різних комбінацій двійкових чисел, а, отже, й знаків). Такі шрифти отримали назву UNICODE. Як і шрифти другого етапу, їх використовують у складі операційних систем на зразок сімейства WINDOWS.

Питання для самоконтролю

1. Які основні завдання стоять перед сучасною прикладною фонетикою?
2. Назвіть основні інструменти для спектрального аналізу звуків.
3. Які критерії застосовують для оцінювання якості графіки:
4. Проаналізуйте основні види систем письма.
5. Які основні завдання стоять перед комп'ютерною лінгвістикою в галузі удосконалення системи письма?

Завдання

1. Визначте оптимальність абеток для української (38 фонем, 33 літери; 36 графем), англійської (46 фонем, 26 літер, 114

графеми) мов.

2. Визначте надлишковість абеток для української (38 фонем, 33 літери; 36 графем), англійської (46 фонем, 26 літер, 114 графеми) мов.
3. Укладіть словник термінів: *абетка, графіка, графема, мовний звук, фонема, орфографія, спеціальне письмо, оптимальність абетки, надлишковість абетки, графема, криптографування, літера, лігатура, ієрогліф, мова жестів, система письма для незрячих, стенографія, транскрибування, транслітерування, штучна мова.*
4. Укладіть електронний словник до теми (20 термінів).
5. Затранслітеруйте та затранскрибуйте будь-яке речення довжиною 10-20 слів.
6. Затранслітеруйте, затранскрибуйте (фонетичною та фонематичною транскрипцією), закодуйте азбукою Морзе речення: *Комп'ютер, як раніше телефон, радіоприймач, магнітофон, диктофон і телевізор, проникає в усі сфери життя сучасної людини.*
7. Визначіть оптимальність та надлишковість абеток для італійської (21 буква, 38 звуків, 43 графеми) та російської мов (33 букви, 42 звуки, 36 графеми).
8. Підготуйте відеолекцію на одне із питань теми.

ТЕМА 4
АНАЛІТИКО-СИНТЕТИЧНЕ ОПРАЦЮВАННЯ ДОКУМЕНТІВ

План

1. Поняття про аналітико-синтетичне опрацювання документів.
2. Індексування.
3. Інформаційний пошук.
4. Реферування.
5. Коректура.
6. Редагування.
7. Переклад.
8. Готування оглядів.
9. Атрибуція текстів.
10. Генерування текстів.
11. Ведення діалогу.
12. Розуміння текстів.

Теоретична частина

Аналітико-синтетичне опрацювання документів – це процеси перетворення інформації, що міститься в первинному документі, з метою створення вторинних документів. У цьому ж сенсі, зважаючи на застосування загальнонаукових методів аналізу й синтезу під час створення інформаційних документів, уживаються й інші терміни, наприклад, «наукова обробка документів», «аналітико-синтетична обробка документів», «інформаційний аналіз/синтез». Основою аналітико-синтетичного опрацювання документів є згортання інформації – зменшення фізичного обсягу інформаційного повідомлення, що поширене в багатьох галузях знання: філософії, природничих науках, інформатиці, бібліографознавстві та ін.

Для задовольняння потреб споживачів застосовуються різні способи згортання інформації: семантичне й лексичне, макро- й мікроаналітичне. Семантичне згортання інформації передбачає зміну інформативності повідомлення, лексичне – перетворює знакову форму повідомлення, залишаючи зміст без змін. При макроаналітичному згортанні наводять формальні відомості про документ і найзагальніші – про його зміст, при мікроаналітичному – детально розкривають зміст документа (основні й побічні теми, аспекти їх розгляду).

Для забезпечення різноманітних інформаційних запитів користувачів існують різні види аналітико-синтетичного опрацювання текстів: складання бібліографічних описів документів, індексування, анотування, реферування, підготовка оглядових творів, науковий переклад, вилучення фактів тощо.

Одним із важливих процесів аналітико-синтетичного опрацювання документів є *індексування*. Згідно з ДСТУ 2395-2000 «Обстеження документа, встановлення його предмета та відбір термінів індексування» визначається, що : «Індексування полягає в ідентифікації змісту документа з метою його наступного відшукання (пошуку). Індексування не передбачає описування документа як фізичної одиниці, хоча деякі параметри щодо форми документа, його вихідних даних тощо, можуть бути занесені до предметного покажчика, якщо інформація такого роду дозволить користувачеві точніше визначити, чи є цей документ релевантним його запиту».

Якість індексування документів базується на:

- вивченні міжнародних та національних стандартів, правил, форматів, методик в галузі каталогізування та пошуку (методологічний блок);
- освоєнні принципів формування машиночитного запису, процесу створення та ведення каталогу та правил створення бібліографічних записів (технологічний блок);
- оволодінні стратегією пошуку та доступу до інформації (інформаційно-пошуковий блок).

Процес індексування документів можна представити у вигляді певної послідовності операцій: перегляд змісту документу; визначення його основної теми та другорядних аспектів, що можуть представляти інтерес для користувачів даної бібліотеки; вибір термінів і формулювання з них рубрик, що відображають виявлені зв'язки, і, кінцевим результатом, стає заміна ключових понять лексичними одиницями *інформаційно-пошукової мови* (*ІПМ*) – класифікаційними індексами, предметними рубриками, дескрипторами.

ІПМ – штучна мова, створена людьми на основі природної мови. Її основне призначення полягає в тому, щоб встановити приналежність того чи іншого документа і запиту до визначеної групи понять, а також, щоб при пошуку по визначеному запиту інформаційно-пошукова система видавала всі документи, що відносяться до тої самої групи понять.

На сьогодні вирізняють три головних принципи індексування:

– *класифікаційний* (на основі інформаційної мови, призначеної для структурного подання документів чи даних за допомогою класифікаційних індексів і відповідних термінів);

– *предметний* (представлення поняття як елемента інформаційної мови чи терміна природної мови засобами предметних рубрик або авторитетних файлів);

– *координатний* (багатоаспектне вираження змісту документа шляхом переліку дескрипторів або ключових слів).

Не встановлюється жодних обмежень щодо кількості термінів індексування. Перелік понять визначається вмістом інформації, що включає документ, з урахуванням завдань індексування. Індексатор повинен бути спроможний ідентифікувати всі поняття документа, що мають потенційну значимість для користувачів тієї чи іншої інформаційної системи. Встановлення будь-яких обмежень щодо кількості понять може призвести до деякої втрати об'єктивності в

процесі індесування і до втрати інформації, яка могла б бути корисною під час пошуку. Для дитячої бібліотеки особливо важливим є надання користувачам необхідної та релевантної інформації, що можливо лише за умови наближення інформаційно-пошукових запитів до природної мови, взаємодії класифікаційного, предметного та координатного принципів розкриття інформації. Індексування проводиться на основі безпосереднього аналізу документа з урахуванням характеру інформаційно-пошукового масиву.

В бібліотечній практиці застосовуються три види індексування: систематизація; предметизація; координатне індексування.

Для систематизації документів у бібліотеках використовують інформаційно-пошукові мови, що відносяться до традиційних бібліотечно-бібліографічних класифікацій. Державний стандарт України визначає *класифікаційну систему* як «інформаційну мову, призначену для структурного подання документів чи даних за допомогою класифікаційних індексів і відповідних термінів і з метою забезпечення реалізації класифікаційного предметного підходу з використанням, у разі необхідності, абеткового покажчика». Найбільш розповсюдженими в Україні є Бібліотечно-бібліографічна класифікація (ББК) та Універсальна десяткова класифікація (УДК).

Бібліотечно-бібліографічної класифікація (ББК) – національна класифікаційна система Росії, що представлена в трьох основних варіантах: повних, середніх і скорочених таблицях, а також розроблених на їхній основі спеціалізованих варіантах і версіях. Науково-дослідний центр ББК має авторське право на видання таблиць і слідкує за його дотриманням.

При машиночитному каталогізуванні значимим видом індексування є предметизація. Термін «*предметизація*» має декілька значень. В професійній сфері бібліотечно-інформаційної діяльності – це предметизація документів, формування словників предметних рубрик та авторитетних файлів предметних рубрик, організація та використання предметних інформаційно-пошукових систем.

Завдання предметизації – з необхідною повнотою та точністю для конкретної інформаційно-пошукової системи представити у вигляді предметних рубрик основний зміст документа та його

форму для забезпечення ефективного інформаційного пошуку.

Об'єктом предметизації може бути окремий документ, його складова частина або сукупність документів.

Предметизація як технологічний процес включає: відбір документів; аналіз змісту та форми документа з метою визначення предмета, аспектів його представлення в каталозі та виявлення зв'язку між ними в тексті; вибір змістовних компонентів (термінів) і формулювання з них рубрик, що відображають виявлення зв'язку; контроль або кінцеве формулювання рубрик за допомогою словників предметних рубрик (авторитетних файлів предметних рубрик або тезаурусів); редагування предметних рубрик (перевірка правильності предметизації); включення предметної рубрики у бібліографічний запис.

Мова предметних рубрик – штучна інформаційно-пошукова мова, що створена на базі природної мови та відповідає вимогам однозначності. Структурною одиницею мови предметних рубрик є предметна рубрика, що призначена для опису змісту та формальних ознак документів або запитів.

Контроль над використанням рубрик здійснюється за допомогою тезаурусів (словників контрольованої мови індексування), авторитетних файлів предметних рубрик та правил методики предметизації.

Тезаурус – контрольований словник термінів із зафіксованими семантичними взаємозв'язками, який охоплює одну чи більше галузей знань. Вступаючи до інформаційного простору сучасного світу, кожна країна повинна потурбуватися про лінгвістичне забезпечення предметного пошуку реалій своєї історії, культурних, інтелектуальних і духовних здобутків.

Крім нормалізованої лексики (індексів УДК і ББК, предметних рубрик і авторитетних файлів) у процесі індексування застосовується і мова ключових слів – координатне індексування. Ключові слова при тематичному розкритті змісту документа використовують в основному для уточнення предметних рубрик: розширюють або звужують їх значення. Словники ключових слів, на відміну від інших інформаційно-пошукових мов, є ненормованими, тому індексування на їх основі ще називають «вільне індексування».

Отже, практичний досвід індексування документів, показує необхідність системного вивчення та використання інформаційно-

пошукових мов. Саме комплекс інформаційно-пошукової мови здійснює суттєвий вплив на формування фонду, каталогізування, організацію внутрібібліотечної технології, безпосередньо впливає на якість довідково-бібліографічного апарату бібліотеки та обслуговування користувачів інформації.

Пошук (search) – це сукупність операцій, які пов'язані з визначенням місцезнаходження об'єктів з заданими характеристиками або ознаками.

Інформаційний пошук або пошук інформації – це пошук неструктурованої інформації, одиницею представлення якої є інформація у довільних форматах.

Завдання інформаційного пошуку стосується пошуку інформації в документах, пошук самих документів, вилучення метаданих з документів, пошуку тексту, зображень, відео і звуку у локальних реляційних і гіпертекстових базах даних.

Термін «інформаційний пошук» був вперше введений Кельвіном Муером (Calvin Mooers) у виступі на конференції у 1950 році. Залежно від ступеня залучення до інформаційного пошуку технічних засобів і участі в ньому людини розрізняють: «ручний», «машинний» і «автоматизований» інформаційний пошук. В автоматизованих інформаційних системах інформаційний пошук забезпечується і здійснюється із залученням лінгвістичних, інформаційних, програмно-технічних, технологічних, організаційних засобів і складених з них комплексів . Безпосередньо інформаційний пошук здійснюється засобами інформаційно-пошукової системи.

Основними критеріями якості результатів інформаційного пошуку є повнота, точність і оперативність пошуку. Інформаційний пошук на практиці це послідовність операцій, спрямованих на надання інформації зацікавленим особам, процес відшукання в деякій множині текстів (документів) всіх тих, які присвячені темі, зазначеній в запиті, або містять потрібні споживачеві факти чи відомості. Предметом пошуку виступає інформаційна потреба користувача, виражена у формі інформаційного запиту. Потреба людей в інформації постійно змінюється, тому її неможливо описати однозначно. Проте інформаційну потребу можна представити у вигляді деякої послідовності її окремих значень, виражених природною мовою у фіксовані моменти часу. Це окреме значення є *інформаційним запитом*, з яким користувач звертається

до системи. Інформаційний пошук повинен забезпечувати вирішення таких основних завдань:

– пошук релевантної інформації, тобто такої, яка відповідає інформаційним потребам;
– пошук аналогічної інформації в суміжних галузях;
– узагальнення та уточнення отриманої інформації;
– аналіз та оцінка інформації, виходячи з власних завдань.

Інформаційний пошук має три основні мети:

1. Пошук необхідних відомостей про джерело і встановлення його наявності в системі інших джерел. Здійснюється шляхом розшукування бібліографічної інформації та бібліографічних посібників (інформаційних видань), спеціально створених для більш ефективного пошуку і використання інформації (літератури, книги).

2. Пошук самих інформаційних джерел (документів і видань), в яких є або може міститися потрібна інформація.

3. Пошук фактичних відомостей, що містяться в інформаційних джерелах, наприклад літературі, книзі тощо. Наприклад, відомості про історичні факти і події, про технічні характеристики машин і процесів, про властивості речовин і матеріалів, про біографічних даних з життя і діяльності письменника, вченого і т.п.

Розрізняють такі види інформаційного пошуку:

Залежно від мети – адресний (формально-механічний) та семантичний (тематичний). *Адресний пошук* – процес пошуку за ознаками, зазначеним у запиті. Для здійснення такого пошуку потрібні наступні умови: 1. Наявність точної адреси документа. 2. Забезпечення певного порядку розташування документів чи веб-сторінок в сховище системи. Адресами документів можуть виступати як адреси веб-серверів і веб-сторінок, так і елементи бібліографічного запису, адреси зберігання документів у сховищі. *Семантичний пошук* – процес пошуку документів за їхнім змістом. Принципова різниця між адресним і семантичним пошуком полягає в тому, що при адресному пошуку документ розглядається як об’єкт з точки зору форми, а при семантичному пошуку – з точки зору змісту. При семантичному пошуку знаходиться безліч документів без вказівки адрес. У цьому принципова відмінність каталогів і картотек.

Від об’єкта пошуку – документальний і фактографічний;

Документальний пошук – процес пошуку в сховищі інформаційно-пошукової системи первинних документів або в базі даних вторинних документів, які відповідають запиту користувача. *Фактографічний пошук* – процес пошуку фактів, що відповідають інформаційному запиту. До фактографічних даними відносяться відомості, отримані з документів, як первинних, так і вторинних, які одержуються безпосередньо з джерел їх виникнення.

Від ступеня використання технічних засобів – ручний або автоматизований. Однак вони тісно взаємопов'язані між собою.

Реферуванням документа називається метод мікроаналітичного згортання інформації з первинного документа наукового або виробничо-практичного професійного характеру, а також сам процес створення реферату як вторинного документа. Реферативна інформація виконує завдання пошуку, оцінки, систематизації, узагальнення й рекомендації фактографічної інформації, вміщеної в первинних документах. Наукові факти – це змістова основа реферату, його принципова відмінність від таких інформаційних документів, як бібліографічний опис та анотація. Головною метою звернення користувача до реферативної інформації є пошук нових знань та ідей, узагальнень і практичних завдань. Відповідно до цього, суть реферування полягає в зіставленні й порівнянні нової інформації з первинного документа з уже засвоєною й використовуваною в суспільній діяльності.

Отже, реферування передбачає: проведення наукового аналізу та оцінки нової соціально значущої інформації (визначення її цінності); рекомендацію соціальної інформації, потрібної користувачам для здійснення певної суспільної діяльності. Реферування передбачає скорочення фізичного обсягу первинного документа із збереженням його основного змісту. Інформацію в процесі реферування ущільнюють, або згортають, у процесі наукового оброблення документа, що пов'язано з його аналізом і відбором найважливіших змістових відомостей: основних положень, фактичних даних, результатів, висновків. Ущільнення інформації, представленої в первинному документі, є інтелектуальним процесом і певним різновидом інтерпретації тексту. Ступінь опанування методики реферування багато в чому залежить від знання законів загальної логіки. Правильне тлумачення первинного документа є неодмінною умовою для референта. Вміння реферувати передбачає також спеціальну

професійну підготовку. Оціночний аналіз змісту реферованого документа, наукова оцінка референтом новизни й корисності інформації неможливі без глибоких спеціальних знань у відповідній галузі науки, техніки, культури тощо. Останнє пояснює вимоги до потреб референта мати відповідні знання про досягнутий рівень у конкретній предметній галузі людської діяльності та про інформаційні потреби визначеного контингенту користувачів реферативних документів. Референт має також чітко уявляти кінцеву мету реферування. Кінцевою метою цього процесу є випереджувальне відображення, передбачення інформаційних потреб споживачів. Референтові слід також знати вимоги, яким має відповідати реферат як вторинний документ.

Реферування починається з читання тексту первинного документа. Практика засвідчує, що референт має володіти трьома видами читання – *ознайомлювальним, вивчаючим і реферативним.* Ознайомлювальне читання спрямовано на загальне ознайомлення зі змістом документа без спеціальної настанови на його дальше сприйняття. Вивчаюче читання – це вдумливе, інтенсивне читання, під час якого відбувається запам’ятовування змістової інформації тексту та її мовних засобів. Реферативне читання полягає в умінні узагальнювати і на цій основі відбирати найсуттєвішу інформацію. Реферування передбачає також правильне розуміння змісту тексту первинного документа. Під розумінням мають на увазі процедуру тлумачення явища, яке вивчається, його інтерпретацію за допомогою системи правил, притаманних прикладній лінгвістиці. Явище буде усвідомленим, якщо знайдено коректні концепції його опису. У розробленні систем автоматизування інформаційних процесів була й залишається проблема аналізу семантичної структури вихідного тексту з метою визначення фактографічної інформації й наступного її узагальнення та синтезу в тексті реферату. Науковці наголошують, що в жодній галузі застосування комп’ютерних технологій немає таких труднощів, як під час виконання завдань семантичного комп’ютерного згортання/розгортання інформації. Труднощі передовсім зумовлені складністю, а інколи неможливістю формалізації й алгоритмізації процесів мислення, що супроводжують різноманітні форми інформаційного аналізу й синтезу. Це одне з найскладніших завдань машинної переробки інформації, оскільки створення реферату як джерела корисної інформації має спиратися на

семантичний аналіз тексту, який поки що не піддається потрібному ступеню формалізації. Формалізація – це представлення внутрішнього змісту повідомлення в зовнішній формі. Зовнішня форма, що належить до рівня явищ, визначається сутністю змісту, тобто внутрішньою формою інформаційного об'єкта. У сфері інформаційних процесів зовнішня, тобто знакова, форма є матеріальною і об'єктивною; внутрішня форма, тобто зміст документа, є ідеальною й суб'єктивною, а об'єктивність вона має лише в тому розумінні, що відображає об'єктивні, інваріантні, загальні зв'язки матеріального світу. За цих умов формалізація інформаційних процесів зводиться до пошуку елементів лексики, граматики, структури, архітектоніки документа, через які можна було б виразити план змісту (семантику) того чи іншого тексту і тим самим обробити (перетворити) семантичну інформацію без звернення до самого змісту первинного документа.

У системі інформаційних комунікацій спостерігається постійне недовикористання накопичених суспільством знань. Причини такого стану полягають насамперед у недосконалості засобів пошуку інформації (не зважаючи на широке впровадження в цю сферу засобів комп'ютерної техніки) і методів аналітико-синтетичної переробки первинного документального потоку. Спеціалістові будь-якої галузі потрібні не документи, а інформація – факти, концепції тощо. Проте інформації дуже багато взагалі і надзвичайно мало зокрема. Роботи в галузі інформаційного аналізу й синтезу спрямовано на усунення цієї суперечності. Їхня кінцева мета – максимальне використання когнітивних (пізнавальних) можливостей первинного документа завдяки машинному виокремленню в ньому самостійних мінімальних релевантних фрагментів, які збираються в поки що гіпотетичні бази знань, звернення до яких сприяло б значною мірою зниженню потреби використання первинного потоку документів. Науковці обговорюють питання формування системи суспільного спостереження за документальним потоком з метою максимального розкриття й використання його ресурсів для виконання завдань науки, техніки, культури. Комп'ютерне реферування є об'єктом науково-технічних пошуків вітчизняних і зарубіжних спеціалістів. Як зазначалося вище, існуючі системи автоматичного реферування не відтворюють сповна повноаспектного змісту першоджерел і процедура зводиться до побудови квазірефератів на основі

статистичного й позиційного аналізу тексту як способу оцінки їхньої інформативності для екстрагування найбільш інформативних його фрагментів. Такий рівень змістової обробки тексту вже не задовольняє зростаючих потреб в інформації, особливо у зв'язку з інформаційними можливостями Інтернету, що акумулює величезні масиви інформації, яку все складніше не лише знайти, а й опрацювати. Саме система автоматичного реферування могла б сприяти користувачеві мережі, даючи йому можливість переглядати на першому етапі відбору потрібної інформації не величезні масиви документів, а короткі й водночас змістовно повноцінні тексти їхніх рефератів. Автоматичне реферування потребує постійного удосконалення, адже саме реферування в недалекому майбутньому, безсумнівно, буде провідним в охопленні широких інформаційних просторів.

Отже, реферування є одним із процесів аналітико-синтетичного перероблення інформації, який спрямовано на оперативне забезпечення фахівців найсуттєвішою фактографічною інформацією з вітчизняних і зарубіжних наукових документів. Світовий і вітчизняний досвід засвідчує, що систему реферативних документів структуровано за галузевим принципом. Утім спеціалісти зазначають, що диференціація науки, з одного боку, та інтеграція наук – з іншого, потребують створення системи централізованого інформування, що дало б змогу враховувати всі світові або національні джерела, збирати відомості, розпорошені в різних первинних документах.

Коректура – це вид аналітико-синтетичного опрацювання текстів документів, у процесі якого перевіряють відповідність копії документа його оригіналові. Якщо в процесі коректури в оригіналі знаходять помилки, то їх також виправляють (в тому числі й у копії), але ця процедура належить вже не до коректури, а до такого виду аналітико-синтетичного опрацювання як редагування.

Потреба в проведенні коректури виникає, наприклад, після передруковування тексту видання в ЗМІ, після його набирання з рукописного оригіналу тощо. Алгоритм проведення цього процесу такий: в копії та оригіналі документа в одній і тій самій позиції порівнюють, чи стоїть там один і той самий знак. Коли знаки не тотожні, це означає, що в копії є спотворення, яке потрібно усунути.

Коректуру за таким алгоритмом (познакове порівняння двох

файлів) в автоматичному режимі виконують сучасні текстові процесори, наприклад згаданий вище Microsoft Word.

Редагування – це вид аналітико-синтетичного опрацювання текстів документів, у процесі якого текст документа приводять у відповідність до існуючих норм, а також здійснюють його творчу оптимізацію. Серед таких норм слід виділити юридичні, етичні, естетичні, політичні, релігійні, композиційні, логічні, лінгвістичні, психолінгвістичні, видавничі й поліграфічні.

Наприклад, у ЗМІ (редакції газети) редагують тексти чергового номера, у книжкових видавництвах – тексти монографій, підручників, енциклопедій, словників, на підприємствах – тексти інструкцій з використання виробів цих підприємств. У процесі опрацювання для покращення тексту редактор може запропонувати авторові змінити якусь частину тексту, видалити її чи, навпаки, додати нову.

У наш час для ручного опрацювання текстів використовують текстові процесори, про які вже йшла мова вище. Поруч із цим все ширше починають використовувати й системи комп'ютерного редагування, функціонування яких базується на визначенні складності тексту, автоматизації контролю орфографії, пунктуації, деяких норм стилістики тощо.

Переклад – це вид аналітико-синтетичного опрацювання текстів документів, у процесі якого заміняють одиниці мови тексту оригіналу на еквівалентні за змістом одиниці іншої мови за умови максимального збереження наявної в тексті оригіналу інформації.

Подамо кілька прикладів. Припустімо, потрібно перекласти англійською речення Оденьте Машу в сарафан з повісті А. С. Пушкіна «Капітанська донька». У принципі, як переклад можна було б запропнувати такі варіанти: Dress Masha in sarafan (при цьому в примітці потрібно було б пояснити, що таке сарафан); Dress Masha in national Russian clothes; Dress Masha in… (а далі вказати такий англійський одяг, який за покроєм набільше схожий на російський сарафан).

Проте всі ці переклади будуть неправильними. Для знаходження правильного потрібно проаналізувати ситуацію, в якій було сказано цю фразу, а тоді стане очевидним, що правильним буде інший переклад: Dress Masha in peasant clothes.

Із середини 20-го століття дослідники-лінгвісти намагаються створити КЛС, призначені для перекладу текстів з однієї мови на

іншу. На сучасному етапі такі системи справді здійснюють такий переклад, проте його результатом є текст, що має суттєві вади (морфологічні, семантичні, синтаксичні тощо).

Проте не треба забувати, що системи комп'ютерного перекладу призначені для перекладу в основному науково-технічних текстів, де вони забезпечують прийнятні результати, особливо враховуючи мінімальну тривалість виконання такого перекладу та його низьку вартість. Крім того, розробники систем комп'ютерного перекладу постійно вдосконалюють свої системи, а тому кількість невдалих перекладів постійно зменшується.

У наш час існує ціла низка систем комп'ютерного перекладу. Одна з них (система комп'ютерного перекладу РУТА-ПЛАЙ) працює в оболонці текстового процесора Microsoft Word.

Хоча це видасться несподіваним, проте останнім часом японські лінгвісти віднайшли ще одну ділянку для застосування систем комп'ютерного перекладу – це системи для перекладу мов тварин. Більше того, ці дослідники не тільки дослідили, а й сконструювали таку систему, яка перекладає близько сотні звуків котів на мову людини. Звуки мови котів перекладають фразами на зразок Дай їсти, Дай пити, Приголуб мене, Відчепись тощо.

Готування оглядів – це вид аналітико-синтетичного опрацювання текстів документів, у процесі якого за певною темою підбирають низку документів, проводять критичну оцінку, систематизацію й узагальнення наявної в них інформації, а далі на цій основі генерують нове повідомлення (огляд), що завершується висновками. Обсяг огляду для науковців, як правило, не обмежують, а для керівних працівників його обсяг не повинен перевищувати 1-2 с.

Огляди на основі аналізу десятків чи сотень документів готують фахівці з великим досвідом роботи, таких фахівців називають аналітиками. Автоматизація цього виду аналітико-синтетичного опрацювання текстів внаслідок її найвищої складності практично не автоматизована.

Сучасна наука виокремлює такі групи оглядів за основними ознаками: 1) Залежно від інтерпретації дійсності: огляди дійсності – публіцистичні огляди, які моделюють картини світу; огляди творів – не лише моделюють картини світу, але і оцінюють елементи моделювання й відображення для пояснення явища або прогресу; 2) За рівнем узагальнення інформації та впливом на аудиторію:

інформаційні – домінує функція повідомлення, панорамне подання сукупності предметів дослідження. У них об'єднані факти, які моделюють панораму подій, що відбуваються в певний період у різних частинах регіону; аналітичні – спонукають до осмислення реалій та вчинків. У них автор не обмежується поданням інформації про те, що сталося в певний періоду в певному регіоні, а створює панораму певної ситуації та зосереджується на ній; 3) За формою відтворення: цілісні; ділені – складаються з окремих частин, блоків, об’єднаних спільною ідеєю. Кожна частина такого огляду може мати окрему назву, бути написана іншим автором; 4) За змістом: універсальні; тематичні – присвячені одній темі (економічні, комерційні, наукові, внутрішні, зовнішньополітичні, міжнародні, кіноогляди, телевізійні, радіоогляди, книжковий, журнальний, газетні, спортивні тощо).

Термін «*атрибуція*» у сучасному мовознавстві не має і дотепер однозначного витлумачення. Так, в Енциклопедії української мови терміни «атрибуція» й «авторизація» визначаються як синоніми: атрибуція – встановлення авторства тексту на основі композиції, способів текстотворення, почерку, мови змісту і позатекстових відомостей про його походження та історію. Атрибуція здійснюється шляхом зіставлення неавторизованого твору з авторизованим, щоб на підставі подібності (відмінності) довести припущення про авторство. Останнє часто є причиною того, що атрибуцію називають ще й авторизацією. В інших словниках ці терміни розрізняють: а) атрибуція – визначення достовірності, аутентичності художнього твору, його автора, місця й часу створення; авторизація – підтвердження авторства, авторського права (Великий тлумачний словник української мови); б) атрибуція – приписування анонімного художнього твору певному авторові; авторизація – надання автором уповноважень, згоди в будь-якій справі (Словник чужомовних слів).

У словнику «Thinkmap visual thesaurus» тлумачення терміна «атрибуція» виходить за межі визначення автора тексту, а як синоніми до нього подано: 1) ascription (приписування певної якості, характеристики людині або предмету); 2) categorization, classification (категоризація, класифікація, групування людей, предметів у подібні класи, категорії); 3) sorting (сортування, упорядкування об’єктів відповідно до певного критерію).

У контексті такої термінологічної невизначеності в мовознавстві сформувалися два підходи до змістового об’єму поняття «атрибуція». Одні науковці вважають, що атрибуція може стосуватися тільки питань визначення автора тексту, інші ж – розглядають її у ширшому розумінні та, окрім визначення автора тексту, застосовують для визначення стилю, тематики, хронологічних особливостей текстів. У рамках цих підходів вирізняють: 1) авторську атрибуцію – встановлення автора тексту (А. Баранов, М. Пещак); 2) неавторську атрибуцію – віднесення тексту до певної мови, стилю, періоду часу, літературної школи, літературного напрямку і т. ін. (С. Бук, П. Вашак, Г. Мартиненко). Завдання як авторської, так і неавторської атрибуції полягає у віднесенні тексту до наперед зафіксованої множини критеріїв групування текстів на основі подібності лінгвостилістичних характеристик (Г. Мартиненко).

Атрибуцію не можна зводити лише до визначення автора тексту, а слід сприймати як розподіл текстів за групами відповідно до певного критерію. Таким критерієм може бути довільна ознака (тематика, стиль, жанр, мова, століття написання тексту тощо) відповідно до поставленого дослідницького завдання. На основі цієї позиції пропонуємо робоче визначення атрибуції як процесу упорядкування довільних текстових документів у групи за функціональним стилем, тематикою або автором за наперед заданими критеріями або визначеними під час цього процесу. Аналіз наукових праць, присвячених вивченню атрибуції текстів різних функціональних стилів, свідчить про різні аспекти здійснення атрибуції художнього стилю, зокрема авторської атрибуції художніх текстів (П. Берков, П. Вашак, Н. Дарчук). Окремий інтерес становить питання вибору оптимальних лінгвістичних параметрів для розмежування текстів за стилем (В. Критська, С. Сушко, О. Шевельов). Останнім часом спостерігається інтерес науковців і до атрибуції електронних повідомлень, листів, передусім дослідники аналізують помилки щодо повторення літер, заміни літер, інверсії літер, пропущення літер, злиття слів (I. Biskub, M. Koopel), структурне оформлення та форматування текстів (K. Calix). Помітною є тенденція і до вивчення атрибуції наукових текстів (А. Коваль, І. Колегаєва, Т. Радзієвська).

Важливою проблемою також є виявлення лінгвістичних

параметрів, придатних для авторської атрибуції текстів. На сьогодні ще не запропоновано усталеного набору параметрів, на основі яких можна було б стандартизувати процеси атрибуції текстів за автором (М. Штокмар, M. Jockers, J. Rudman). Розроблено класифікацію оптимальних параметрів для здійснення атрибуції текстів, які скласифіковано на мовні (синтаксичні, лексичні, морфологічні, знаки пунктуації, помилки) та позамовні (структурні критерії).

Генерування (породження, синтез) тексту – це процес створення тексту, що включає: виділення фрагмента внутрішнього подання, який увійде у текст; формування схеми викладення інформації; заповнення схеми мовними виразами.

Виділення фрагмента внутрішнього подання, що увійде у текст, та формування схеми викладення інформації є складною лінгвістичною проблемою і передбачає створення та використання моделі тексту. Наприклад, математична експлікація згаданої лінгвістичної проблеми дає змогу структурні моделі будови одиниць тексту на рівні його синтаксису моделювати у вигляді графів (ієрархічних) предикативних чи непредикативних залежностей між словами і словосполученнями.

Природним є змоделювати процес генерації тексту на синтаксичному рівні за допомогою графа, що відтворює паралельні процеси, тому є математичною абстракцією для моделювання побудови тексту.

Природна мова є доступним і універсальним засобом ведення діалогу, а у поєднанні із мультимедійними засобами (звук, колір, графіка) дозволяє безперешкодно здійснювати комунікацію.

Діалог – це вербальна інтеракція, що передбачає обмін вербальними висловлюваннями, вимовленими двома людьми або людиною і машиною.

Ефективність використання усного мовлення для ведення діалогу вражає: середньостатистична людина здатна за хвилину надрукувати близько 20 слів, написати 24 слова або вимовити 150 слів. Мовлення, без сумніву, є найприроднішим способом комунікації.

Існують принципові відмінності між письмовою та усною комунікацією. Під час написання автор має можливість розмірковувати над формулюванням речення, він може модифікувати його до досягнення бажаного результату. Читач

потенційно наділений здатністю перечитувати написане за умови виникнення непорозумінь або сумнівів. Під час генерування звукового мовлення помилки можуть бути виправлені, але не вилучені. Процедура корекції відбувається у режимі реального часу і супроводжується повторами, парантезами, змінами просодичних характеристик тощо. Усний діалог між людиною і машиною обмежений технологічними можливостями сучасних систем розпізнавання мовлення, а також складною структурою і варіативністю звукового потоку, згенерованого людиною. Основні проблеми виникають через спонтанність продукування висловлювань, які часто містять надмірну лінгвістичну інформацію, повтори, самовиправлення, скорочення, порушення правил граматики й синтаксису. Існує також потенційна можливість перехоплення ініціативи в діалозі, припинення комунікації через різноманітні причини тощо. Під час мовленнєвого діалогу з комп'ютером розв'язання цих проблем відбувається, переважно, шляхом уведення додаткових класифікаційних субдіалогів, моделювання яких є окремим напрямом у сучасних автоматичних мовленнєвих технологіях.

Порівняно з діалогом, що ведеться між людьми, комунікація з комп'ютером характеризується чітким плануванням комунікативних кроків, майже повною відсутністю переривань.

Здатність діалогових систем до метамислення є предметом інтенсивного наукового пошуку, однак уже зараз цілком очевидно, що когнітивний потенціал природної мови дає змогу вибудовувати комунікативно-когнітивні моделі поведінки комп'ютера під час мовленнєвого спілкування з людиною і відображати результати такого моделювання у зручній для користувача формі, у вигляді повідомлень природною мовою. Генерування мовленнєвого діалогу між користувачем і комп'ютером вимагає формалізованого кодування і подальшої автоматичної обробки різноманітних типів знань.

Розуміння – необхідний елемент будь-якого комунікативного акту, коли комунікант x (учасник дискурсу) дещо стверджує (пише), а комунікант y (інші учасники дискурсу) слухають або читають з метою засвоєння й осягнення того, що сказав або написав комунікант x. Розуміння смислу промов і текстів постає в історичному контексті як особливий соціокультурний феномен у структурі «культура – мислення – мова», яке досліджують

інформатика, семіотика, герменевтика, логіка.

Кожний текст як дещо цілісне структурують на такі частини: інформація в тексті; мова, якою зображена інформація; смисл інформації, що закодована в тексті, який треба зрозуміти. Текст вводиться в поле розумово-мовленнєвої діяльності людини в процесі його засвоєння.

Засвоєння тексту структурується на рівні: лінгвістичному (засвоєння лексики, якою зображено текст); історичному (стиль мовлення, який визначає історичний час створення тексту та його особливості); евристичному (наявність нового знання (інформації); логічному (логічність міркувань, наявність послідовності).

Засвоєння тексту і його розуміння – не одне і те саме, тобто розуміння не зводиться лише до запам'ятовування або точного переказу тексту, хоча без засвоєння немає розуміння. Розуміння – такий рівень засвоєння знання (інформації) в тексті суб'єктом х, коли він осягає смисл прочитаного і виражає його в певному понятті. Поняття в контексті розуміння означає концепцію (сутність), яка міститься в тексті. Розуміння смислу означає, що суб'єкт х на підставі логічного препарування тексту визначає для себе сутність того, що він прочитав, виразив сутність у певному понятті й встановив істинність тексту як дещо цілісного. Розуміння як феномен виникає у процесі засвоєння тексту, коли існує діалог автора тексту та суб'єкта який читає текст, визначає для себе його значущість і раціональними засобами намагається осягнути смисл тексту.

Питання для самоконтролю

1. Що таке аналітико-синтетичне опрацювання документів?
2. Назвіть різновиди аналітико-синтетичного опрацювання документів.
3. Який вид аналітико-синтетичного опрацювання текстів через його найвищу складність практично неможливо автоматизувати?
4. Який вид аналітико-синтетичного опрацювання документів передбачає творчу оптимізацію?
5. Який вид аналітико-синтетичного опрацювання документів передбачає ідентифікацію змісту документа з метою його подальшого пошуку?
6. Що є основою аналітико-синтетичного опрацювання документів?

7. У чому специфіка машинного реферування й анотування тексту шляхом його стиснення?

8. Що таке надфразові зв’язки у тексті.

9. Опишіть методику тезаурусного реферування науково-технічного тексту.

10. Назвіть об’єктивні оцінки якості реферування.

11. Опишіть процедуру створення еталонного реферату.

12. У чому специфіка реферування на основі пофразної семантичної мережі?

13. Назвіть основні види реферування, анотування тексту.

14. Назвіть два підходи до створення реферату, анотації.

15. Проаналізуйте основні методи, на яких базується лінгвістичний аналіз побудови реферату, анотації.

Завдання

1. Укладіть словник термінів: *аналітико-синтетичне опрацювання документів, індексування, інформаційний пошук, реферування, коректура, редагування, переклад, готування оглядів, атрибуція текстів, генерування текстів, ведення діалогу, розуміння текстів.*

2. За допомогою програми Pragma здійсніть комп’ютерний переклад тексту російською та англійською мовою. Зробіть кількісний (статистичний) та якісний аналіз помилок у перекладних текстах:

Був собі один чоловік, мав шестеро синів та одну дочку. Пішли вони в поле орати і наказали, щоб сестра винесла їм обід. Вона каже.

– А де ж ви будете орати? Я не знаю. Вони кажуть.

– Ми будемо тягти скибу від дому аж до тії ниви, де будемо орати, – то ти за тією борозною і йди.

Поїхали.

А змій жив над тим полем у лісі та взяв тую скибу закотив, а свою витяг до своїх палаців. От вона як понесла братам обідати та пішла за тією скибою і доти йшла, аж поки зайшла до змієвого двора. Там її змій і вхопив.

Поприходили сини ввечері додому та й кажуть матері.

– Ввесь день орали, а ви нам не прислали й пообідати.

– Як-то не прислала? Адже Оленка понесла, та й досі її нема. Я думала, вона з вами вернеться. Чи не заблукалась?

Брати й кажуть.

– Треба йти шукати її.

Та й пішли всі шість за тією скибою і зайшли таки до того змієвого двора, де їх сестра була. Приходять туди, коли вона там.

– Братики мої милі, де ж я вас подіну, як змій прилетить? Він же вас поїсть!

Коли це й змій летить.

– А, – каже, – людський дух пахне. А що, хлопці, биться чи миритьсяся?!

– Ні, – кажуть, – биться!

– Ходіть же на залізний тік!

Пішли на залізний тік биться. Не довго й бились, як ударив їх змій, так і загнав у той тік. Забрав їх тоді тільки живих та й закинув до глибокої темниці.

3. Придумайте список ключових слів для інформаційного пошуку веб-ресурсів, присвячених огляду проблеми аналітико-синтетичного опрацювання документів.

4. Створіть мультимедійну презентацію (20 слайдів).

5. Здійсніть інформаційний бібліографічний пошук за планом практичного заняття.

6. Застосувавши будь-яку традиційну систему автоматичного реферування та анотування, здійсніть автоматичне реферування статті Міщенко А. Л. Сучасні методи, напрямки й здобутки комп'ютерної лінгвістики / А. Л. Міщенко // Наукові записки Кіровоградського державного педагогічного університету імені В.Винниченка. – Вип. 95 (2). – Серія: Філологічні науки (мовознавство). – Кіровоград: РВВ КДПУ ім. В.Винниченка, 2011 . – С. 122–233

7. Використовуючи запропоновані сервіси (або інші доступні сервіси), виконайте автореферування текстів різних стилів. Проаналізуйте одержані результати. Автореферат тексту якого стилю виконано краще? Чому? Який із сервісів виконує автореферування краще?

http://visualworld.ru/referat.jsp
http://textcompactor.com/
http://about.viwo.ru/referat.html

Параметр аналізу	**Коментар**
Зв'язний текст або набір речень	
Функційна навантажені елементів автореферату	
Чи відображені основні структурні елементи,	

тема, мета?	
Загальний висновок	

ТЕМА 5
ПРОБЛЕМИ СТВОРЕННЯ ШТУЧНОГО ІНТЕЛЕКТУ

План

1. Інтелект як інструмент пізнання дійсності.
2. Складники інтелектуальної діяльності людини.
3. Складові штучного інтелекту.
4. Підходи до створення систем штучного інтелекту.
5. Машина та тест Тьюринга.
6. Визначення можливості створення штучного інтелекту.
7. Місце лінгвістичного забезпечення в системах штучного інтелекту загального призначення.
8. Складність створення систем штучного інтелекту.

Теоретична частина

Людина, на відміну від інших представників тваринного світу, має здатність мислити. У науковій літературі мислення ототожнюється з поняттям інтелекту. *Інтелект* – це здатність людини пізнавати світ і вирішувати проблеми, що об'єднують в собі пізнавальні здібності. Виходячи з цього поняття, можна говорити про те, що інтелект – це риса, необхідна для когнітивної діяльності людини.

Існує два підходи у вивчені інтелекту як інструменту пізнання дійсності:

1. *Генетичний*, що полягає у представленні інтелекту як суто спадкової риси. Швидкість сприйняття навколишнього середовища не може залежати від людини, адже вона або народжується зі здатністю до високих інтелектуальних здібностей, або ні.

2. *Структурний,* що полягає в здатності людини швидко реагувати на зовнішні фактори та сприймати їх.

Кібернетика розглядає мислення людини як інформаційний процес, фіксує те загальне, що є в роботі електронно-обчислювальних машин і в мисленні людини. А психологію насамперед цікавить специфіка людського мислення, його відмінності від інформаційних процесів, які реалізуються через сучасні технічні пристрої. Взаємодія психології та «штучного

інтелекту» суттєво змінила зміст психології мислення.

Отже, існує безліч поглядів на поняття інтелекту, але засадою виокремлення цієї здібності у людини є свідома діяльність індивіда, формування його когнітивних та інших властивостей. Рівень розвитку інтелекту визначає, якою мірою людина здатна орієнтуватися в навколишньому середовищі, як вона опановує обставини і себе.

Інтелект реалізується за допомогою інших людських здібностей. Це, наприклад, здатність пізнавати світ, навчатися, логічно мислити, систематизувати інформацію шляхом її аналізу, класифікувати інформацію, знаходити у ній зв'язки, закономірності та відмінності тощо.

На інтелектуальний розвиток особистості великою мірою впливають соціальні, етнокультурні, психоментальні чинники. Для рівня розвитку інтелекту людини, як засвідчують дослідження нейропсихологів, філософів, когнітологів, вирішальними виявилися такі ознаки мисленнєвої діяльності:

1. Повнота та адекватність сприйняття дійсності.

2. Уміння правильно ставити завдання розумової діяльності, добирати потрібну й всебічну інформацію про об'єкти дійсності, структурувати певну предметну галузь – об'єкт розумової діяльності, а саме: знаходити її релевантні ознаки і встановлювати зв'язки між ними.

3. Відповідно до норм та уявлень певного суспільства оцінювати, категоризувати й класифікувати одержану інформацію про певну предметну галузь; робити несуперечливі й адекватні висновки та узагальнення.

4. Знаходити рішення, які дозволяли б якомога повніше й успішніше розв'язувати поставлені завдання й досягати мети розумової діяльності у найпростіший спосіб. Якість процесу реалізації інтелекту характеризують критерії, які також визначають рівень інтелекту людини:

1. Швидкість здійснення необхідних процедур.

2. Вибір оптимального способу виконання завдання, найраціональнішого шляху до досягнення поставленої мети.

3. Уміння добирати й організовувати необхідні знання про досліджувані об'єкти з належною для виконання певного завдання повнотою та всебічністю.

4. Здатність будувати адекватну та несуперечливу модель

об’єкта або процесу з високим ступенем пояснювальної та передбачувальної сили, тобто модель, придатну для ефективного опису, аналізу та пояснення.

Складники інтелектуальної діяльності людини:

1. Пізнання (набуття знань про навколишній світ);
2. Розуміння (розкриття внутрішньої суті предметів, явищ і процесів);
3. Зберігання (закріплення знань та їх використання);
4. Генерування знань (вміння використовувати знання про ту чи іншу сферу життя в конкретних ситуаціях);

У 1956 році, під час роботи двомісячної літньої школи з проблем комп’ютерного оброблення інформації, була організована конференція в Дартмутському коледжі (США). Підсумком роботи конференції стало введення в науковий обіг терміна «штучний інтелект». Цей термін стоїть на межі як мінімум трьох галузей знань: кібернетики, інформатики та комп’ютерної лінгвістики. Тому цей термін має декілька дефініцій. У межах проблем цих дисциплін виокремлюють дві основні дефініції терміна.

Штучний інтелект – 1) наука й технологія, що займається вивченням і створенням інтелектуальних машин і комп’ютерних програм; 2) здатність інтелектуальних систем виконувати мисленнєві процеси, що є прерогативою людини.

На сьогодні виділяють декілька складових штучного інтелекту: доведення теорем; розпізнавання образів (слухових і зорових); теорія ігор; адаптивне, динамічне й евристичне програмування; прийняття рішень; природна мова та її машинне розуміння, або спілкування з ЕОМ природною мовою; системи з самоорганізацією та саморегулюванням, синергетика; роботика; створення комп’ютером музики; самонавчальні мережі; оброблення даних, представлених природною мовою; вербальне й концептуальне навчання.

Існують різні методи створення систем штучного інтелекту. У наш час можна виокремити чотири досить різні методи:

1. *Логічний підхід*.

Основою для вивчення логічного підходу слугує алгебра логіки. Кожна машина, що створюється на основі логічного підходу, має блок генерації мети, і система виводу намагається довести дану мету як теорему. Якщо мета досягнута, то послідовність використаних правил дозволить отримати ланцюжок

дій, необхідних для реалізації поставленої мети (таку систему ще називають експертною системою).

2. Структурним підхід

Це спроба побудови штучного інтелекту шляхом моделювання структури людського мозку. Тут головною моделюючою структурною одиницею є нейрон. Пізніше виникли й інші моделі, відоміші під назвою нейронні мережі.

3. Еволюційний підхід

Під час побудови системи штучного інтелекту за даним методом основну увагу зосереджують на побудові початкової моделі, і правилам, за якими вона може змінюватися (еволюціонувати). Модель може бути створена за найрізноманітнішими методами, це можуть бути нейронні мережі, набір логічних правил і будь-яка інша модель. Після цього комп'ютер, на основі перевірки моделей, відбирає найкращі з них, і за цими моделями за найрізноманітнішими правилами генеруються нові моделі. Серед еволюційних алгоритмів класичним вважається генетичний алгоритм.

4. Імітаційний підхід

Об'єкт, поведінка такого штучного інтелекту є «чорним ящиком». Не важливо, які моделі в нього всередині і як він функціонує, головне, щоб модель в аналогічних ситуаціях поводила себе без змін. Таким чином тут моделюється інша властивість людини – здатність копіювати те, що роблять інші, без поділу на елементарні операції і формального опису дій. Часто ця властивість економить багато часу об'єктові, особливо на початку його життя.

Для узагальненого представлення процесу комп'ютерного моделювання розумової діяльності людини англійський математик Алан Тьюринг у 1936 р. запропонував у своїй статті «Про обчислювані числа» гіпотетичний пристрій – прообраз комп'ютера, який міг читати, тобто сприймати, розуміти, або розпізнавати мовну інформацію у графічній формі, а також писати й стирати символи, тобто приймати рішення, представляти їх у певній графічній формі і змінювати одне прийняте рішення на інше. Роботу такого уявного пристрою, який згодом дістав назву «машина Тьюринга», можна було описати найпростішим алгоритмом. На кожному кроці роботи такого алгоритму дія «машини Тьюринга» визначена її поточним станом, або тим, що вона уміє робити, «знає» на цей момент, фігурально

висловлюючись, «рівнем її розумового розвитку, інтелекту», та символом, який вона зчитує на цьому кроці роботи, тобто надходженням нового завдання для її «розумових здібностей, інтелекту». Так Алан Тьюринг у найпростішому вигляді змоделював процес розумової діяльності людини. Алану Тьюрингу належить і знамените формулювання проблеми моделювання людського інтелекту – «Чи може машина мислити?», відповідь на яке й шукають розробники систем штучного інтелекту.

У 1950 р. відомий математик А. Тьюринг, висловлюючись з приводу можливості створення штучного інтелекту, заявив, що штучний інтелект буде створено тоді, коли, розмовляючи з машиною (через якісь опосередковані засоби, наприклад, канал передачі інформації), людина не зможе відрізнити цю систему від людини, тобто людина не знатиме, з ким розмовляє. Отже, саме спілкування визначатиме, створено штучний інтелект чи ні.

У системах штучного інтелекту, які призначені, наприклад, для конструювання будівель, автомобілів, керування водопостачання, складання холодильників чи пральних машин тощо, функцію спілкування людини з самою системою виконуватиме лінгвістичне забезпечення, або лінгвістичний інтерфейс.

Основне завдання прикладних і комп'ютерних лінгвістів – разом з іншими дослідниками – взяти участь у створенні такого штучного інтелекту, який би міг спілкуватися з користувачами природною мовою та опрацьовувати тексти з такою ж ефективністю, як людина. Проте поки що наука досконало не знає механізмів мислення людини, методів її творчості, механізмів мовного спілкування. Тому створення штучного інтелекту в наш час перебуває на стадії дослідження й експериментів.

Принцип розвитку систем штучного інтелекту є таким: чим більше штучний інтелект наближається до людського, тим його створення стає важчим. Аналогічну ситуацію спостерігаємо при створенні комп'ютерних систем, що супроводжуються такою закономірністю: якщо кількість команд у системі збільшити в n разів, то складність її налагодження зросте в n^2 разів. Отже, коли ми маємо якусь програму завдовжки 200 рядків, а далі збільшимо її довжину вдвічі (тобто доведемо до 400 рядків), то складність її конструювання й налагодження зросте у 22, тобто в 4 рази.

Потреба розв'язання таких задач вимагає зіставлення їх

складності з можливістю алгоритмізації та програмування. Лише після такого зіставлення, а також аналізу ресурсів (кадрових, фінансових тощо) можна приймати рішення про доцільність створення певної комп'ютерної лінгвістичної системи.

Питання для самоконтролю

1. Назвіть основні досягнення в історії дослідження проблеми створення штучного інтелекту.
2. Охарактеризуйте напрямки досліджень в галузі штучного інтелекту.
3. В чому полягала критика штучного розуму?
4. В чому можливості та переваги штучного інтелекту?
5. Яка роль біологічного фактора у створенні штучного інтелекту?
6. Які існують проблеми у створенні систем штучного інтелекту?
7. В чому полягають філософські аспекти проблеми штучного інтелекту?
8. Що таке експертні системи?
9. Які функції виконують експертні системи?
10. В чому специфіка мови штучного інтелекту?

Завдання

1. Укладіть словник термінів: *інтелект, інтелект природний, інтелект штучний, коефіцієнт інтелектуальності, «машина Тьюринга», «тест Тьюринга».*
2. Підготуйте реферат на одну із тем:

Експертні системи в (медицині, економіці, екології, логістиці, спорті, сільському господарстві, оподаткуванні, міжнародних відносинах, інформатиці, лінгвістиці).

ТЕМА 6
АВТОМАТИЧНИЙ СИНТАКСИЧНИЙ АНАЛІЗ

План

1. Передумови досліджень в галузі автоматичного синтаксичного аналізу (АСА).
2. Особливості здійснення процедури АСА.
3. Текстові процесори.
4. Системи редагування.
5. Технологічні особливості комп'ютерного редагування.

Теоретична частина

1. Кожен із нас щодня будує десятки, сотні речень. Як це вміння передати машині? Як навчити машину розуміти синтаксичну структуру речення, а також будувати нові правильні речення? Зрозуміло, що наше знання про синтаксичну структуру речення, тобто про лексико-граматичні зв’язки слів у ньому передати ЕОМ неможливо, оскільки в алгоритми аналізу речення не можна ввести команди типу «знайди підмет», «знайди прикметник, який визначає іменник» тощо, якщо немає детальних, автоматично виконуваних правил про те, як це робити. Адже машина розуміє тільки мову команд, а не мову їхнього розв'язання. Для того, щоб ці правила створити, необхідно пізнати ті закони, які діють у процесі побудови речення.

Існує два підходи щодо дослідження цього процесу. Мову можна уявити у вигляді кібернетичної системи, на вході якої є сума речень, а на виході – класи мовних одиниць і правила їх сполучуваності. Або навпаки: на вході системи – породжувані цією системою речення. Ці два підходи пізнання структури мови лежать в основі побудови синтаксичних моделей (індуктивних та дедуктивних) та розробляються у методиці структурних лінгвістичних досліджень.

Метод моделювання змусив переглянути існуючі синтаксичні теорії, а також точніше визначити основні поняття синтаксису, розробляти нові методи його вивчення. Заново були поставлені основні проблеми синтаксису: проблема його об'єкта, співвідношення із семантикою й морфологією; проблема слова, групи, фрази як синтаксичних одиниць, а також проблема основних понять синтаксису: відношення (зв’язку), функції, структури, формальних показників.

Чимало цікавих ідей, використаних для розробки автоматичного синтаксичного аналізу (АСА), висловили представники дескриптивної школи структурної лінгвістики: із суми спостережень над текстом лінгвіст здобуває первісне уявлення про спосіб організації тексту й у вигляді чітких процедур – правил алгоритму – повідомляє автомату свої дії, а потім за його допомогою одержує на більшому матеріалі дані, що цікавлять дослідника.

У роботах з АСА прийнято два способи опису синтаксичної

структури:

1) опис за безпосередніми складниками (БС);

2) опис за допомогою дерев залежностей, які називають деревами синтаксичного підпорядкування.

Ці два способи допомагають описати синтаксичну структуру на двох рівнях:

а) за допомогою БС описуються в явному вигляді словосполучення, але не розпізнається «хазяїн» і «слуга»;

б) дерева залежностей дають можливість розрізнити характер зв'язків між словами. Якщо в результаті роботи алгоритму АСА встановлюються зв'зки, які більш-менш відповідають інтуїтивним уявленням носіїв мови, значить синтаксична структура речення «визначена» правильно.

Завдання АСА полягає у тому, щоб, використовуючи морфологічну інформацію про словоформи, одержану на попередньому морфологічному етапі, побудувати синтаксичну структуру вхідного речення. Об'єктом аналізу є речення, яке до моменту синтаксичного аналізу подається у вигляді інформаційних ланцюжків до словоформ. Виконувати синтаксичний аналіз повинен алгоритм СА, тобто інструкція, яка складається зі стандартних елементів, що здійснюють певну послідовність операцій над словоформами. Результатом аналізу є синтаксична структура речення, представлена як сукупність даних про синтаксичні зв'язки між його одиницями.

В основі АСА лежить формально-синтаксичний аспект вивчення речення. Ані семантико-синтаксичний і функціональний, ані комунікативний підхід до розгляду речення не можуть стати основою автоматизації. Тоді як дослідження формально-синтаксичної будови речення дає можливість створити словник синтаксем, для якого попередньо слід укласти таксономічну класифікацію лексики, що у майбутньому уможливить автоматичне визначення синтаксичних відношень між членами словосполучення. Формальна граматика, адаптована для потреб автоматизації, базуватиметься на гіпотаксисі (синтаксичний зв'язок підпорядкування одного компонента (повнозначного слова, сполучення слів) іншому в словосполученні чи реченні (зв'язок узгодження, керування, прилягання) як провідному аспекті синтаксичного ладу мови; а паратаксис (сурядність) буде додатковим аспектом, оскільки виокремлення сурядних

словосполучень з погляду автоматизації не становить суттєвих труднощів.

У ході автоматичного синтаксичного аналізу речення насамперед має здійснюватися автоматичний пошук зв'язків слів у реченні. Ознаки таких зв'язків наявні, зокрема, у словозмінних характеристиках слів. У реченні послідовно розгортається підпорядкування слів одне одному: одне слово (залежне) змінює форму, щоб адаптуватися до вимог іншого слова (головного). Таким чином, машина має виокремлювати пари слів, пов'язані граматичним зв'язком, позначаючи напрямок залежності.

Наприклад, для речення: *Широко (1) обговорюються (2) проблеми (3) життя (4) українського (5) суспільства (6)*. – виокремлюються такі пари слів:

(5) ← (6) українського суспільства
(1) ← (2) широко обговорюються
(3) → (4) проблеми життя
(4) → (6) життя суспільства
(2) ↔ (3) обговорюються проблеми.

Цим діям можна надати алгоритмічного вигляду. Врешті отримуємо список пар залежностей. Жодних даних семантичного характеру у цьому аналізі не використовується. Єдине, що можна визначити при такому підході, – це залежність слів одне від одного і порядок їх розташування. Це і є прикладом формально-синтаксичного підходу до аналізу речення. Словосполучення – це смислове та граматичне поєднання двох або більшої кількості слів на основі підрядного, сурядного або предикативного зв'язку. Ці типи зв'язків відповідають відтворенню загальної системи відношень між компонентами описуваної ситуації у реченні. Віднесення до словосполучень тільки тих, які сполучаються підрядним прислівним зв'язком, не є вичерпним з точки зору складників речення. Ми вважатимемо, що словосполучення – відносно самостійна одиниця мови, що виділяється у межах речення, будується за законами поєднання слів, виявляє у мовленні валентні властивості головного слова, має мовні моделі, відтворювані у мовленні. Не є словосполученнями: складені аналітичні поєднання слів, зокрема сполуки іменника з прийменником (через міст, в інституті); складені аналітичні форми слів (буду читати, більш досвідчений); фразеологізми (ні пари з вуст, бити байдики).

Завданням АСА є виявлення всіх різновидів сполучуваності – предикативної, підрядної і сурядної – кожного слова з текстів. Граматичні характеристики словосполучення безпосередньо залежать від того, до якої частини мови належить слово-«хазяїн», тому що лексико-граматична природа слова визначає його здатність сполучатися з іншими словами. Відповідно до цього словосполучення поділяють на іменникові, прикметникові, займенникові, числівникові, дієслівні та прислівникові.

За виробленою концепцією АСА при виокремленні словосполучень було передбачено попередній етап створення словника валентностей для дієслова (31206 правил), іменника (40023), ад'єктива (6205), а також словника фразеологізмів (близько 3000 одиниць). За складом словосполучення поділяють на прості, складні та комбіновані. Ми виділяємо тільки прості бінарні словосполучення, які можуть бути поширені у складні або комбіновані автоматизовано, оскільки при визначенні їх складу потрібен аналіз смислової структури. Якщо у сполуках у головній позиції слово інформативно недостатнє, а залежне слово цю недостатність заповнює (так звані доповнювальні, або комплетивні відношення), то вони розглядаються як словосполучення, що виконують функцію одного члена речення, наприклад: дехто з присутніх, четверо з них, почав працювати і под. Ми відмовляємося від традиційного поділу підрядного зв'язку на підвиди – узгодження, керування та прилягання. Ґрунтуючись на широкому розумінні поняття синтаксичного зв'язку, всі вони є випадками приєднання до головного слова відмінкової форми іменника або субстантива. При узгодженні залежне слово уподібнюється до головного в усіх його граматичних формах, а при приляганні воно, не маючи форм словозміни, приєднується до головного за змістом. Крім того, останнім часом у деяких роботах випадки поєднання з головним словом залежної форми іменника з атрибутивним чи обставинним значенням трактуються як прилягання: *працювати лікарем* – зв'язок керування, а *прогулюватися парком* – прилягання; *допомога матері* – керування. Причиною такої відмови є ще й неможливість у деяких випадках автоматично, не вдаючись до аналізу значення кожного з членів словосполучення, визначити його тип. Алгоритм АСА має спиратися виключно на морфологічні форми слів (орудний відмінок залежного слова у першій парі; давальний у другій).

Підрядні зв'язки поділяються нами на ядрові і неядрові. Ядровим називаємо такий зв'язок, при якому аналізоване слово є керувальним, головним. Наприклад, у реченні – *Від економічної кризи сильно постраждали майже всі європейські держави*. – спостерігаємо такі ядрові зв'язки: **кризи** домінує над економічної; **постраждали** домінує над від; **від** домінує над кризи; **постраждали** домінує над сильно; **всі** домінує над майже; **держави** домінує над всі; **держави** домінує над європейські. Неядровий зв'язок – це такий зв'язок, при якому аналізоване слово є залежним, керованим. У попередньому прикладі неядрові зв'язки є у словах економічної (залежить від кризи), сильно (залежить від постраждали), європейські (залежить від держави) тощо. Предикативний зв'язок – це зв'язок між основними компонентами речення «підмет – присудок», який ґрунтується на їхній двобічній залежності. У предикативній парі жодне зі слів не можна вважати домінованим, вони обидва є однаково домінувальними. Якщо підмет або присудок виражений складеним словосполученням, то визначається підрядний зв'язок для аналізованого слова, наприклад: *Двоє→студентів почали→скандувати*. Предикативний зв'язок установлюємо між двоє студентів і почали скандувати. В межах словосполучення двоє студентів ядровим буде двоє, яке домінує над студентів (студентів, відповідно, має неядровий зв'язок); у словосполученні почали скандувати ядровим буде почали, яке домінує над скандувати.

Те саме стосується іменного складеного присудка:

Він став→студентом. Ядровий зв'язок встановлюємо між підметом він і присудком став студентом. У межах іменного складеного присудка став студентом ядровим буде допоміжне дієслово став, тому що воно домінує надіменною частиною, вираженою іменником в орудному відмінку студентом. Сурядний зв'язок – це зв'язок, при якому жодне із взаємопов'язаних слів не є ані домінувальним, ані домінованим. Вважається, що два слова знаходяться в сурядному зв'язку, якщо кожне з них підпорядковане одному й тому ж третьому слову, якщо вони пов'язані через сполучник між собою або відокремлені одне від одного комою. При цьому ми дотримуємося таких умов, що сурядний зв'язок установлюється між словами, а не між словом і зворотом або синтаксичною конструкцією.

Наприклад, сурядний зв'язок є між словами *чесними і*

прозорими (*Вибори були чесними і прозорими*) і його немає у такому прикладі: *Президент був задоволений і в гуморі*. Щодо ад'єктивів (прикметників, дієприкметників, займенників), які виконують одну й ту саму функцію, прийняті такі домовленості: якщо між ними є сполучник, то напрямок зв'язку такий: від головного слова до кожного з прикметників, а потім прикметники із сполучником, наприклад: *порядні і достойні←банкіри* (порядні банкіри, достойні банкіри, порядні і достойні); якщо між ними безсполучниковий зв'язок, то сурядні зв'язки встановлюються з кожним із ад'єктивів та іменником (у будь-якій формі), а потім між самими ад'єктивами, наприклад: *порядні, достойні банкіри* (порядні банкіри, достойні банкіри, порядні, достойні). Кожний тип словосполучення відображається у певному виді моделі. Модель словосполучення – це двоелементна формула, що відбиває один із типів зв'язку аналізованого слова з певним повнозначним словом, наприклад: прикметник + іменник (видатний діяч); іменник + іменник (коло друзів); дієслово + прислівник (працював важко). У тих випадках, коли прийменник (або сполучник) служить лише засобом зв'язку між двома повнозначними словами, він не вважається самостійним членом моделі. Таким чином, модель «дієслово + прийменник + іменник» (працювати в уряді) залишається двочленною, хоча складається з трьох слів. Подаються чотири типи моделей: ядрові; неядрові (ад'юнктні, які відображають підрядні зв'язки); сурядні; предикативні.

АСА здійснюється за правилами – моделями, представленими у таблиці SyntaxRules у програмному середовищі Access, яка виконує роль диспетчера. Кожній моделі згідно з таблицею автоматично приписується певний код. Зокрема, за цією таблицею здійснюється «збирання» в один вузол складених морфологічних та синтаксичних явищ, наприклад: ГБ – аналітичний майбутній час (буду читати); ГЗ – аналітичний наказовий спосіб (хай читає); ГЧ – умовний спосіб (читав би); СЧ – складений числівник (сорок три); ПМ – складений підмет (один з них); ПС – складений присудок (почав працювати). Простим присудкам і безособовим формам дієслова приписуються коди: ПР та ГЧ відповідно. Причому в першому випадку далі ведеться пошук підмета, а в другому пошук продовжується за таблицею дієслівної валентності. Типи синтаксичних словосполучень кодуються за частиномовною належністю: ІС – іменникове; АС – прикметникове; ДС – дієслівне;

ЧС – числівникове; РС – прислівникове; ЗС – займенникове. За цією ж таблицею кодуються види синтаксичних зв’язків (vpr): КЗ – координація; ПЯ – підрядний ядровий; ПА – підрядний ад’юнктний; СУ – сурядний. Напрямки перевірки (праворуч / ліворуч) строго регламентуються набором правил для конкретного рядка, якими визначається і пріоритет у роботі групи правил.

Як свідчить тестування результатів роботи АСА, автоматично виділялися такі словосполучення і такі типи зв’язків, які повністю відповідають інтуїтивним уявленням носіїв української мови та експертів-лінгвістів. Отже, синтаксичну структуру речення автомат «зрозумів» правильно. І це відкриває широкі перспективи, зокрема можливість укладання частотного словника сполучуваностей української мови та здійснення автоматичного синтаксичного аналізу цілого речення. У свою чергу, правильний синтаксичний аналіз є запорукою створення автоматичного семантичного аналізу тексту.

Сьогодні існують текстові процесори, у яких крім операцій набирання й виправлення тексту, автоматизовано найпростіші операції редагування. Завдяки низьким цінам ТП доступні для будь-кого з користувачів. Редагування україномовних тестів поки що має дуже низький ступінь автоматизації.

Текстові процесори (ТП) – це програми, які дають змогу набирати, виправляти й зберігати текст, а також виділяти й формувати компоненти його видавничої структури (рядки, абзаци, сторінки, розділи тощо). Фактично, такі текстові процесори є системами редагування із найнижчим ступенем автоматизації (це – комп’ютеризовані системи редагування).

До основних функцій текстових процесорів (набирання й виправлення тексту, комп’ютеризація редагування) раніше та й зараз додають функції поліграфічних систем. У результаті такі ТП стають змішаними (гібридними) редагувально-поліграфічними системами. Останнім часом до ТП дедалі частіше додають також функції власне редагування (контроль лінгвістичних і психолінгвістичних норм – автоматичну перевірку синтаксичної зв’язності слів, знаків пунктуації, наявності пасивних синтаксичних конструкцій тощо). Виправлення після такого контролю здійснюють у комп'ютеризованому чи автоматизованому режимах. Крім того, до найпотужніших ТП для полегшення роботи редакторів часто включають ще й різні лінгвістичні словники

(наприклад, синонімів та антонімів).

Поряд із автоматизацією контролю лінгвістичних і психолінгвістичних норм у ТП включають процедури, що дають змогу автоматизувати конструювання видання, забезпечуючи при цьому дотримання видавничих норм. Сюди належать: автоматичне нумерування і перенумерування будь-яких компонентів видання (рубрик, таблиць тощо); автоматичну зміну нумерації в посиланнях при зміні номера компонента видання; автоматичне укладання змісту видання із зазначенням номерів сторінок, на яких розташовані рубрики; автоматичне укладання покажчиків видання із зазначенням номерів сторінок, на яких розташовані ключові слова; автоматичне розташування й нумерування колонтитулів; будування розділів і підрозділів повідомлення у форматі ієрархічної структури.

Сучасні ТП мають також функції, що дають змогу набирати складні тексти (формули й таблиці), а також виконувати технічні малюнки середньої складності. Останні версії ТП мають також конвектори (програми, що змінюють формат подання даних у файлах), які дають змогу перетворення видання, підготовані для друкування на папері, у комп'ютерні видання (тут мається на увазі перетворення поліграфічних команд у команди формату HTML), доступні через мережу Інтернет.

До ТП, які в найбільшому обсязі автоматизують функції редагування, без сумніву, належить текстовий процесор Microsoft Word – класичний приклад гібридної редагувально-поліграфічної системи (на відміну від таких НПС, як Corel Ventura, Page Maker, Microsoft Publisher, TEX та ін.).

Враховуючи сучасні можливості комп'ютерної лінгвістики, можна передбачити, що розвиток систем редагування і далі буде здійснюватися на основі текстових процесорів. Проте цілком імовірно, що функції редагування і поліграфічного конструювання з часом у них будуть все більше розділятися і на базі текстових процесорів виникнуть спеціалізовані окремо функціонуючі системи редагування (СР). Саме ці СР забезпечуватимуть можливість автоматизованого й автоматичного редагування текстів.

Імовірно, що при цьому паралельно будуть функціонувати й гібридні редагувально-поліграфічні системи.

4. Якщо розглядати редагування як переклад неправильної мови на правильну, системи редагування доцільно порівнювати із

системами комп’ютерного перекладу (СКП), а також із системами транслювання програм (трансляторами) (через здійснення формально-логічного та синтаксичного контролю; відмінність: транслятори призначені для опрацювання штучних, а СР – природних мов). СР із СКП об’єднує те, що вони не тільки перекладають із природних мов, а й виконують також морфологічний, синтаксичний, а іноді ще й семантичний аналіз (відмінність, звичайно, полягає в тому, що СКП не здійснюють адаптацію повідомлень до певних умов). Незважаючи на схожість з СКП, СР специфічні, оскільки редагування має цілу низку своїх завдань, зокрема оцінювання, які у вказаних системах не реалізовані. Редагування – це не переклад, а приведення повідомлення у відповідність із нормами. Значно більше аналогій у СР з експертними системами (ЕС). ЕС об’єднують можливості комп’ютера зі знаннями та досвідом експерта в такій формі, що система може запропонувати розумну пораду чи здійснити розумне вирішення поставленої задачі.

Для виконання таких функцій ЕС повинні: мати знання в певній галузі (постійні знання); мати знання про особливості кожного конкретного випадку (змінні знання); на основі знань та конкретних даних про досліджуваний об'єкт вміти робити висновки (давати поради чи оцінки); вміти пояснювати свої рекомендації. У цьому плані ЕС є різновидом систем штучного інтелекту. Будь-яка конструкція типу «порівняй дві величини і виконай дію, що залежить від результату порівняння» може бути використана в ЕС. ЕС функціонують на опрацюванні тверджень. Чим більше тверджень є в базі знань ЕС, тим вищий її інтелектуальний рівень. Постійними знаннями для СР, збудованої у формі ЕС, можуть служити наявні норми редагування, вставлені в конструкції умовних тверджень, а змінними – дані про читацьке призначення, жанрові й стилістичні особливості повідомлення. Така система зможе давати розумне пояснення пропонованих чи виконаних нею процедур виправлення. Отже, СР належать до експертних систем. Таке визнання має не тільки теоретичне, а й практичне значення. Воно, зокрема, дає підстави для побудови ефективних систем редагування, які – на відміну від людини – не хворіють, не забувають норм редагування, контролюють повідомлення на основі об'єктивних, а не суб'єктивних норм і не потребують витрат на оплату праці. Прикладом такої СР можуть служити деякі функції

редагування в текстовому процесорові Microsoft Word, де операції синтаксичного контролю англомовних текстів реалізовані за принципами ЕС (з можливістю подання потрібних пояснень виконаних дій). Але повністю ототожнювати СР з ЕС не можна.

Таким чином, 1) СР найефективніше конструювати за принципами побудови ЕС; 2) СР мають деякі відмінності від ЕС, які полягають у тому, що на СР покладають не лише вироблення рекомендацій, а й їх реалізацію (в цьому розумінні СР є складнішими за ЕС); 3) як і ЕС, СР належать до систем штучного інтелекту.

СР як окремо сконструйованих систем ще дуже мало (найчастіше це програми перевірки орфографії). Значно поширенішими \ текстові процесори, у яких, крім перевірки орфографії, автоматизовано цілу низку інших функцій редагування. Але з часом кількість спеціально сконструйованих СР як окремих продуктів імовірно зростатиме.

Завдяки комп'ютерному редагуванню процес коректури або суттєво скорочується, або стає зовсім зайвим.

Якщо традиційно спершу проводять редагування, а далі – коректуру, то комп'ютеризована технологія, коли подають авторський оригінал на папері, вимагає спершу проводити коректуру й тільки потім – редагування. Адже в текст недоцільно вносити виправлення, коли заздалегідь відомо про наявність у ньому спотворень.

Комп'ютерне редагування має особливості, які відрізняють його від традиційного. Зокрема, в ньому можна виділити такі ступені автоматизації: комп'ютеризоване редагування (операції контролю і виправлення здійснює людина; комп'ютер використовують лише як «електронне перо» ; прикладом СР, які дають змогу здійснювати комп'ютерне редагування, є ТП); автоматизоване редагування (більшу частину операцій контролю виконує СР, а людина – меншу частину операцій контролю та більшість операцій виправлення); автоматичне редагування (більшу частину операцій контролю й виправлення виконує СР, а меншу – людина; крім того, людина приймає рішення в неформалізованих конфліктних ситуаціях).

Якщо під час традиційного редагування людина опрацьовує текст інтегрально, тобто охоплює одночасно кілька сусідніх рівнів, то комп'ютерне редагування здійснюється послідовно: від однієї

одиниці до іншої, переходити до вищого рівня можна лише тоді, коли на нижчому всі помилки вже усунуті.

Межі автоматизації редакційного етапу

На 100 % автоматизувати редакційний етап неможливо, оскільки навіть людина-редактор не може усунути з повідомлення ні практично, ні теоретично всіх помилок. У наш час ступінь редагованості, який забезпечують СР, є значно нижчим, ніж у людей-редакторів.

Розвиток СР стримують такі чинники:

1) відсутність формалізованих повних переліків норм редагування (формалізації тяжко піддаються норми, що мають форму положень, особливо ті, що стосуються смислового аспекту редагування повідомлення.

2) неавтоматизовано семантичне опрацювання повідомлення.

Отже, на сучасному етапі бар'єром у розвитку СР є семантика повідомлень.

У наш час, коли СР ще не можуть замінити людину-редактора, їх слід використовувати як помічників (асистентів) людей-редакторів. У цьому зараз полягає основна функція СР.

Лінгвістичні норми

У наш час, коли в пристрої оперативного запам'ятовування можна в повному обсязі записати весь орфографічний словник будь-якої мови, застосовують лише словникові методи комп'ютерного редагування – які базуються на тому, що контролюють не окремі частини слів, а відразу цілі слова. При цьому використовують не автокореляційні, а кроскореляційні методи, оскільки вони видають для перевірки меншу кількість слів (автокореляційні пропонують у середньому в два-чотири рази більшу кількість). Серед помилок, що залишаються після опрацювання тексту за допомогою цього методу контролю, граматичні спотворення становлять 27%, полілексемні – 20, пунктуаційні – 16, семантичні – 11, поліграфічні (наприклад, спотворення в шрифті) – 8, а спотворення в цифрових і змішаних текстах – 18 %. Суть кроскореляційного словникового контролю полягає в тому, що найперше в пам'ять комп'ютера записують словник потрібної мови. Далі кожне слово контрольованого тексту шукають у комп'ютерному словнику і, якщо знаходять, то вважають його правильним, а якщо ні – помилковим. Звичайно, і серед правильних слів можуть бути помилкові (у випадку, коли

замість одного правильного слова ввели зовсім інше правильне слово, тобто допустили семантичну помилку), а серед помилкових можуть бути правильні (у випадку, коли такого слова – вузькогалузевого терміна, неологізму тощо – в словнику нема).

Обсяг комп'ютерних словників для контролю тексту в більшості діючих програм перевірки орфографії перебуває в межах кількох десятків тисяч слів. Різке збільшення обсягу цих словників до сотень тисяч веде лише до зниження ефективності контролю (наприклад, частина неправильних слів через наявність у словнику абревіатур ідентифікуються як правильні).

Зрозуміло, що не існує такого комп'ютерного словника, в якому були б усі слова, які використовують у тексті. Адже постійно з'являються нові терміни, неологізми, нові абревіатури. Тому вважають дуже доброю ситуацією, коли покриття тексту комп'ютерним словником становить близько 98 %.

Щоб забезпечити максимальне покриття тексту, яке істотно залежить від семантичного наповнення словника, часто чинять так: у комп'ютерному словнику виділяють словник загальновживаної лексики (загальний словник) і лексику окремих галузей знань (галузеві словники). У кожний галузевий словник, крім термінів і номіналій, включають також персоналії (власні імена і прізвища людей), географічні назви, скорочення та абревіатури. Далі для контролю тексту якоїсь конкретної галузі використовують загальний словник, а також добирають потрібний галузевий словник. Крім того, деякі програми перевірки орфографії дають користувачам змогу створювати для окремих видань локальні комп'ютерні словники, тобто словники для одного конкретного тексту.

Комп'ютерні словники, які використовують для редагування текстів, класифікують: за типом лексичних одиниць – словники словоформ, в яких слова подають у всіх їх словозмінних формах, і словники основ, у яких до основи кожного слова вказують всі її можливі закінчення; за наявністю блоків для аналізу морфем (префіксів, суфіксів і закінчень), за допомогою яких від основ можна утворювати нові похідні слова.

Велике значення під час контролю тексту має швидкість пошуку слів. Для її підвищення використовують різні способи організації слів у пам'яті комп'ютера. Наприклад, сортують слова за алфавітом, за довжинами або використовують спеціальні функції

кодування для прямого виходу відразу на потрібне слово. Достатньою вважають швидкість пошуку близько 100 слів за секунду (для аналітичних мов) і 30 слів за секунду – для синтетичних мов.

Для аналітичних мов (на зразок англійської) частіше використовують словники словоформ, а для синтетичних (як українська) – словники основ. Кожен із цих словників має свої переваги та недоліки. Так, словник словоформ дуже легко можна створити, опрацювавши на комп'ютері достатньо великий масив текстів і записавши всі однакові слова у вигляді словника на комп'ютерний носій інформації. На жаль, такий словник за обсягом у кілька разів буде перевищувати аналогічний словник основ, який можна створити лише традиційним способом. Тому інколи для синтетичних мов використовують комбінований тип словників.

Методи реконструкції. Операції виправлення значно складніші, ніж операції контролю. Вони дають змогу автоматично виправляти лише окремі знаки в словах. Серед цих методів найвідоміші абревіатурний, алфавітний, базовий, комбінаторний та цифровий.

Крім описаних автоматизованих методів реконструкції, в деяких ТП використовують і метод повністю автоматичної реконструкції, який користувач за своїм бажанням може вмикати чи вимикати. Такий метод полягає в тому, що помилки в деяких часто вживаних словах є однаковими у великої кількості людей, а тому можна задати їх автоматичне виправлення за допомогою реконструюючого словника підстановок (наприклад, завжди заміняти зпід на з-під, свойого на свого, твойого на твого тощо). За допомогою такого словника можна контролювати й автоматично заміняти суржик і типові часто повторювані помилки.

Для контролю орфографічної правильності україномовних текстів зараз на ринку програмних продуктів наявні системи RUTA, PLAJ I Language Master (остання система, крім перевірки орфографії, дає змогу реалізувати автоматичний переклад текстів для трьох мов). Ці системи здійснюють словниковий кроскореляційний контроль. Вони мають у своєму складі орфографічний словник сучасної української літературної мови, працюють на IBM-сумісних комп'ютерах в операційній системі Windows і, як правило, розраховані на перевірку орфографії у власному текстовому процесорі або паралельне використання з

певним текстовим процесором. Існують також програми перевірки орфографії україномовних текстів для Macintosh комп'ютерів. Крім контролю, ці системи виконують також автоматизовану реконструкцію, тобто пропонують редакторові на вибір кілька варіантів виправлення помилкового слова. Вкрай потрібним у наш час є доповнення україномовних СР тлумачними словниками, словниками синонімів й антонімів. Комбінування програм перевірки орфографії з лексикографічним інструментарієм збільшить цінність таких програм.

Для синтетичних мов, до яких належить і українська, важливим є проведення контролю синтаксичної зв'язності слів у реченнях. Такий контроль дає змогу знаходити в тексті помилки на зразок з веселий сміхом. Його суть полягає в перевірці всіх означень, що стоять у препозиції до іменника, на наявність синтаксичного узгодження в роді, числі та відмінку з означуваним іменником. Такий контроль без особливих труднощів може бути реалізований за допомогою комп'ютерного словника української мови, який дає змогу здійснювати морфологічний аналіз тексту. Для інших синтетичних мов (наприклад, російської) операції такого контролю вже функціонують. На рівні речень актуальним є також контроль правильності синтаксичного підпорядкування слів. Проведення такого автоматичного синтаксичного контролю в СР є достатньо складним, але цілком можливим.

Відповідні програми контролю можуть бути розробленими на основі моделей порядку слів у реченні.

Перспективним для рівня синтаксем є також контроль таких лінгвістичних помилок: тавтологій (наявність у фрагменті повідомлення великої кількості варіантів слова з одним і тим самим коренем, тобто низький ступінь урізноманітнення тексту для образних та образно-понятійних повідомлень); подвійних заперечень (психолінгвістичними дослідженнями встановлено, що будь-який елемент тексту із запереченням є складнішим, ніж той самий елемент без заперечення); речень у пасивному стані (психолінгвістами встановлено, що будь-яке речення в пасивному стані є складнішим, ніж в активному).

Для деяких європейських мов (наприклад, англійської, російської) вже створені й функціонують програми перевірки пунктуації. В Україні є лише експериментальні системи перевірки пунктуації. Розгляньмо методи контролю, використовувані у цих

системах.

У першій зі СР для перевірки правильності пунктуації використовують автоматичний морфологічний та синтаксичний аналіз тексту. При цьому контролюють лише дієприкметникові звороти. Незважаючи на високу ефективність контролю окремих зворотів, загальна ефективність контролю всіх розділових знаків у тексті в цій системі, звичайно, низька.

У другій СР (експериментальній), що названа «Редактор», контролю, як і в традиційному редагуванні, підлягають лише обов'язкові розділові знаки. Для контролю використовують індикаторний метод, який базується на тому, що існує ціла низка розділових знаків, для контролю яких достатньо виявити в тексті лише певні елементи (індикатори), котрі однозначно вказують на необхідність розділового знака (наприклад, сполучники а, але в середині речення однозначно вимагають коми). Ці індикатори названі формальними. Крім них, існують ще й такі, що вимагають попереднього автоматичного морфологічного, синтаксичного та семантичного аналізу. Оскільки не всі із цих видів аналізу під силу сучасним СР, то контролюють у системі «Редактор» лише формальні індикатори. За межами контролю залишаються однорідні члени речення, дієприкметникові та прикметникові звороти тощо.

Виходячи з імовірнісного характеру індикаторного принципу, СР «Редактор» не розставляє розділових знаків сама, а лише вказує на місця, де можливі помилки. Правильність розділових знаків, поставлених людиною, для СР є остаточною і контролю не підлягає.

СР працює за таким основним алгоритмом: якщо в реченні є пунктуаційний індикатор, біля якого не стоїть потрібний розділовий знак, то система повідомляє користувачеві, що тут помилконебезпечне місце і надає інформацію, потрібну для виправлення.

Для англійської та російської мов у ТП Microsoft Word реалізовані деякі достатньо прості види стилістичного контролю. Для україномовних текстів частина функцій стилістичного контролю реалізована в уже згадуваній експериментальній СР «Редактор». У СР «Редактор» для програмної реалізації були вибрані лише ті найпростіші норми, які не пов'язані з контролем семантики, а саме: контроль прийменників, сполучників, префіксів

та часток, що впливають на милозвучність мови (наприклад, чергування прийменників у-в, сполучників і-й-та, часток ся-сь тощо). Задача стилістичного контролю отримала таку цільову функцію: на границях повнозначних слів так модифікувати текст повідомлення, щоби кількість відкритих складів у ньому була максимальною. Обмеження полягало в тому, що на границях повнозначних слів під час виправлення збіг однакових звуків та складів є забороненим.

Для формування бази даних індикатори стилістичних ситуацій відбирали з урахуванням їх частотності, для чого було використано «Частотний словник сучасної української художньої прози». У самій базі даних записи впорядковували за ймовірністю їх появи в текстах повідомлень.

СР «Редактор» здійснює стилістичний контроль у діалоговому режимі: найперше вона запитує користувача про стиль, до якого належить повідомлення (користувач повинен вибрати один із чотирьох стилів: розмовно-побутовий, публіцистичний, художній, науковий). Далі СР виявляє у тексті відхилення від норми і пропонує користувачеві готові варіанти виправлень, які користувач може приймати, або відкидати, заміняючи своїми. При бажанні користувач може отримати пояснення пропонованого виправлення.

Очевидно, що чергування в стилістиці – це аж ніяк не єдина група норм, контроль яких можна автоматизувати. Контролювати можна й показник багатства словника автора, тобто відношення кількості різних слів тексту до їх загальної кількості.

Контроль за дотриманням психолінгвістичних норм почали застосовувати після широкого впровадження персональних комп'ютерів, на яких автори за допомогою ТП набирали, а редактори виправляли тексти повідомлень. ТП виявилися тим інструментом, який дав змогу легко підраховувати й виражати в кількісній формі значення параметрів деяких психолінгвістичних норм, повідомляючи авторів про необхідність виправлення рукопису відповідно до планованої реципієнтської аудиторії

Для англомовних текстів відповідність повідомлень психолінгвістичним нормам визначають кілька ТП, зокрема Microsoft Word та Word Perfect, для російської мови – русифікована версія Microsoft Word, а для україномовних текстів – експериментальна СР «Редактор».

Труднощі тут виникають при визначенні кінців речень, а

також при визначенні границь абзаців. Робота з СР «Редактор» передбачає, що в систему найперше завантажують текст, який повинен підлягати контролю. Визначення складності тексту відбувається в діалоговому режимі. Передусім користувачеві пропонують повідомити, для якої реципієнтської аудиторії він підготував своє повідомлення (Вкажіть складність тексту: дуже легкий, легкий, не дуже легкий, звичайний, не дуже важкий, важкий, дуже важкий).

Ефективність визначення семантичної складності повідомлення (відношення кількості слів повідомлення, що відсутні в усередненому словнику реципієнтської аудиторії, до їх загальної кількості в повідомленні) істотно залежить від якості укладеного словника-мінімуму. Такий словник повинен бути укладений на базі якомога більшої кількості частотних словників. Для української мови було використано, на жаль, лише два таких словники – художньої прози та публіцистики.

Питання для самоконтролю

1. Що передбачає правильна синтаксична структура?
2. Що таке автоматичний синтаксичний аналіз (АСА)?
3. Яка роль дистрибутивного методу і методу безпосередніх складників (БС) у розробці (АСА)?
4. Які лінгвістичні передумови граматики залежності?
5. В чому специфіка граматики залежностей як способу представлення синтаксичної структури речення?
6. Як зображуються структури з однорідністю?
7. Що таке дерева залежностей у системах АСА?
8. Проаналізуйте принципи роботи деяких синтаксичних аналізаторів: системи АОТ, Dictum, аналізатор на основі системи SMART.
9. Що таке синтаксична омонімія?

Завдання

1. За допомогою сервісу http://slashzone.ru/parser/ зробіть розбір речень: «*Он достал трубку из глины»; «Он надел пальто на меху»; «Он надел пальто на улице»*. Чи демонструють результати двозначність речення 1? Чи пропонує сервіс альтернативний варіант синтаксичного розбору? Чому? Скільки синтаксичних моделей запропоновано сервісом для 2 і 3 речень? Чому?

2. Наведіть приклади 5 українських, російських та англійських речень, де можлива синтаксична омонімія. Як вирішується проблема синтаксичноїомонімії цих речень сучасними сервісами АСА.
3. Запропонуйте аудіословник до теми.
4. Порівняйте специфіку роботи сервісів, визначте їх основні переваги та недоліки:
   http://nlp.stanford.edu:8080/parser/index.jsp
   https://www.connexor.com/nlplib/?q=demo/syntax
   http://www.link.cs.cmu.edu/link/submit-sentence-4.html
   http://slashzone.ru/parser/

ТЕМА 7
АВТОМАТИЧНИЙ СЕМАНТИЧНИЙ АНАЛІЗ

План

1. Напрямки формалізації семантики.
2. Автоматичний логіко-семантичний аналіз тексту:

- методики визначення в тексті ключових слів (слів-концептів);
- автоматичне індексування тексту;
- різновиди систем інформаційного пошуку (ІПС);
- інформаційно-пошукові мови (ІПМ): класифікаторні та дескрипторні;
- інформаційно-пошукові тезауруси (ІПТ).

Теоретична частина

Автоматичний семантичний аналіз (АСЕА) є однією з найактуальніших і разом з тим найскладніших проблем комп'ютерної лінгвістики, оскільки пов'язаний із проблемами моделювання людського інтелекту. Під **автоматичним семантичним аналізом** звичайно розуміють сукупність методів і прийомів, за допомогою яких можна шляхом однозначних формальних процедур за правилами певної формалізованої граматики, що реалізується на комп'ютері за спеціальними лінгвістичними алгоритмами, з досить високою точністю однозначно представити смисл.

Виникає питання, який ми, власне, чекаємо від комп'ютера семантичний аналіз тексту? Наприклад, на вході системи є текст. Перед комп'ютером можуть стояти такі завдання, пов'язані зі

смисловим аналізом: 1) зробити переклад іноземною мовою; 2) «розуміти» питання щодо теми тексту і давати на них відповіді; 3) зробити стислий реферат тексту; 4) «зрозуміти» тему тексту та автоматично проіндексувати його за темою; 5) зробити атрибуцію тексту тощо. Що об'єднує і що розмежовує ці завдання у плані їх розв'язання? Для всіх проектів зі створення штучного інтелекту, пов'язаних із лінгвістичним забезпеченням автоматичних систем опрацювання науково-технічної інформації (машинний переклад, діалогові системи, анотування, реферування, атрибуція тексту, автоматична індексація тощо), повинна будуватися універсальна база, яка стає єдиною основою для різних систем автоматичного опрацювання інформації. Універсальність, яка є найважливішою її характеристикою, визначається тим, що при всіх видах опрацювання тексту (МП, анотуванню тексту, діалозі «машина-людина» тощо) потрібна граматична і лексична інформація, яка створюється й організується так, щоб її можна було використовувати в різних системах АОТ без перебудови самої бази. Вона повинна мати оптимальну організацію відносно завдань, які можуть бути поставлені перед різними системами АОТ. Оптимізація лінгвістичного забезпечення передбачає модульність його структури у вигляді рівневої модульної системи, яка складатиметься з таких трьох компонентів:

1) лінгвістичне програмне забезпечення – програмний комплекс із морфологічного, синтаксичного, семантичного анотування;

2) словникове забезпечення – лінгвістично-інформаційна база;

3) базове програмне забезпечення.

Рівні аналізу визначаються типом робочої моделі як сприйняття, так і генерації тексту, послідовністю планованих процедур і необхідним інформаційним словниковим забезпеченням, яке також повинно будуватися на принципах рівневого підходу, відкритості. На основі єдиного базового програмного комплексу сумісність дозволить компонувати систему, виконувати конкретні завдання аналізу і синтезу, які заздалегідь визначаються поставленою метою АОТ. При цьому відкритість є головним фактором ефективного функціонування АОТ: уведення нової інформації не потребує перебудови всієї системи. Отже, словникове забезпечення є інформаційною базою будь-якого АСЕА і повинно бути реалізовано у вигляді ієрархічної,

модульносумісної і відкритої системи.

Не можна ігнорувати смисловий зміст текстуальних відношень, що важливо при передаванні смислу мовних висловлювань. Значимість слова проявляється у тісних зв'язках з мережею смислових відношень, які пов'язують слово з іншими елементами не тільки речення, а й абзацу чи цілого тексту. Якщо в ідеалі було б можливим створити певну формальну граматику, а в її межах розробити мову смислу, тоді можна було б розв'язати проблему моделювання людського мислення. Або навпаки, навчившись моделювати людське мислення, можна створити формальну граматику, розробивши правила здобуття смислу з тексту. На жаль, упродовж майже 60 років активної роботи над проблемами штучного інтелекту якихось усталених універсальних методів аналізу змісту тексту не виявлено, тому що формалізація семантики, яка є необхідною умовою вміння виражати смисл, є надважким завданням. Це пов'язано з тим, що при моделюванні смислу тексту треба виходити за межі мови і звертатися до зовнішнього світу, до класифікації предметів, уявлень, які знаходяться поза межами мови. Основою розуміння тексту як людиною, так і комп'ютером виступає попереднє формування уявлень про зовнішній світ на основі аналізу інформації, яка міститься у зв'язному тексті. Результатом такого аналізу є виокремлення головних компонентів ситуації, визначення їх диференційних ознак і включення до системи категорій людського досвіду, що можна умовно вважати розумінням тексту. Зусилля найкращих лінгвістів і цілих колективів в усьому світі спрямовані на подолання труднощів перетворення формальної структури тексту в його смислову форму, вироблення теоретичних засад цього процесу.

Перший напрямок – це дослідження, які проводяться на дедуктивному абстрактно-теоретичному рівні, мета яких – встановити співвідношення між семантикою і семіотикою, з одного боку, і синтактикою та прагматикою, з іншого; побудувати моделі людського мислення взагалі й у зв'язку з процесом комунікації; вивести універсальні закономірності утворення понять, зв'язку між поняттями і значеннями слів, висловлюваннями та їхніми складовими; врешті, встановити зв'язок між мисленням і комунікацією.

Другий напрямок має індуктивний емпіричний характер, мета

його – розв'язання конкретних прикладних завдань: машинного перекладу, автоматичного інформаційного пошуку, автоматичного реферування, анотування, індексування. І якщо для другого напрямку характерним є уникання інформаційних моделей зі складною «граматикою», то для першого – навпаки, створення такої «граматики» є необхідністю: складність завдань корелює зі складністю апарату дослідження. Існують два теоретичні напрямки семантичного аналізу речення з уведенням граматики: інтерпретаційна граматика Ч. Філлмора і породжувальна граматика Н. Хомського. Прибічники інтерпретаційної теорії семантики вважають, що смисловим представленням речення є інтерпретована фразова структура, термінальні вузли якої заповнюють лексичні одиниці. Згідно з цією теорією інтерпретація глибинної структури (з елементами семантичного аналізу) здійснюється семантичним компонентом, який передбачає наявність спеціального словника, де кожне слово супроводжується смисловим та семантичним маркером, а також вказівкою на обмеження для даного слова, тобто сумісної сполучуваності слів. Синтаксичні маркери мають вигляд категорійних символів типу NP (група підмета), VP (група присудка), S (символ речення) тощо. **Семантичні маркери** – це смислові диференційні ознаки типу істота / неістота, рахується / не рахується тощо. Ці маркери відрізняють одні значення від інших. Селекційні обмеження мають характер указівок на семантичні маркери, які мають бути обов'язково присутніми в елементах, що сполучаються з даним словом. Особливе місце в межах інтерпретаційної семантики посідає робота Ч. Філлмора, який вважає, що основу речення створює предикатно-аргументна структура, де аргументами є імена, для яких указано глибинний відмінок. Під терміном **відмінок** розуміється глибинно-семантичне відношення, а під терміном **відмінкова форма** – вираження відмінкового відношення у конкретній мові. Речення у своїй глибинній основі трактується як таке, що складається з дієслова й однієї чи більше іменних груп, кожна з яких зв'язана з дієсловом певним відмінковим відношенням. Смисли відмінків утворюють набір універсальних понять, що ідентифікують певні типи суджень, які людина здатна робити про події, що відбуваються навколо неї, судження про речі, такі, як: Хто зробив щось? З ким щось трапилося? ... Відмінок у Філлмора розглядається як універсальне явище, властиве всім мовам – це узагальнене відношення між

дієсловом і змістом однієї з його іменних груп.

Усього пропонується сім глибинних відмінків:

- агентивний (A) (відмінок живого ініціатора дії, ідентифікованого з дієсловом, напр.: *Джон відчинив двері; Двері були відчинені Джоном*);
- інструктивний (I) (відмінок неживого предмета або сили, що є причиною дієслівної дії або стану, напр.: *Ключ відчинив двері; Джон відчинив двері ключем; Джон скористався ключем, щоб відчинити двері*);
- давальний (D) (відмінок живої істоти, пов'язаної з дієслівною дією, напр.: *Джон вірив, що він виграє; Джону було очевидно те, що він виграє*);
- фактитивний (F) (відмінок предмета або дії, що виникає внаслідок дії, або входить як частина самої дієслівної дії, напр.: *Вітрина розбилася*);
- локативний (L) (місце або просторова орієнтація дієслівної дії, стану, напр.: *Чикаго вітряний; У Чикаго вітряно*);
- бенефактивний (B) (відмінок користувача) (*читач задоволений*);
- об'єктивний (O) (семантично найбільш нейтральний відмінок, відмінок чогось, що може бути позначено іменником, роль якого відносно дії визначається семантичною інтерпретацією самого дієслова, напр.: *Джон відчинив двері*).

Склад відмінків та характеристики не є остаточними. Глибинна структура за Ч. Філлмором має такий вигляд: S = A + P, де S – речення, A – модальний показник, P – пропозиція (смисловий елемент текстових структур).

Теорією, в якій немає принципової різниці між синтаксисом і семантикою, є теорія породжувальної граматики Н. Хомського. Завдання породжувальної граматики – розкриття внутрішніх закономірностей мовної структури, яка представляється у вигляді певного «механізму», що піддається не тільки спостереженню у процесі природного функціонування мови, а й приведенню в дію штучним шляхом. Для цього вихідний матеріал треба представити у вигляді правильних речень (моделей) і надати кожному з них один або кілька структурних описів. Граматика будується як дедуктивна система, що передбачає раціональний метод пояснення всієї реальної різноманітності функцій слів і, відповідно,

синтаксичних побудов. Тоді мовленнєвий процес описується не в термінах, наприклад, безпосередніх складників, а в термінах упорядкованого набору правил, необхідних для побудови речення, що моделює його породження. Граматика мови розглядається Н. Хомським як система правил, що ставить у зв'язок кожному правилу значення кожного породжуваного нею речення. Під **породжувальною граматикою** учений розуміє систему правил, яка експліцитно надає реченням структурні описи. Очевидно, що кожен мовець володіє породжувальною граматикою, яка відображає його знання мови, своєю чергою, знання мови має імпліцитну здатність розуміти необмежену кількість речень, тому породжувальна граматика повинна бути системою правил, за якими можна породжувати нескінченно велику кількість структур (із різними компонентами – фонологічними, синтаксичними, семантичними).

Різниця між породжувальною та інтерпретаційною семантикою починається з розуміння смислової структури слова. Якщо в інтерпретаційній граматиці значення слова складається із семантично невпорядкованих наборів елементарних смислів, то в межах породжувальної граматики елементарні смисли синтаксично впорядковані, тобто зосереджені не в одному якомусь вузлі фразової структури (як це має місце в інтерпретаційній семантиці), а розподілені по різних вузлах цієї структури, які перебувають в ієрархічних відношеннях. Можна сперечатися з приводу ролі і місця інтерпретаційної граматики Ч. Філлмора і породжувальної граматики Н. Хомського в системах з автоматичного семантичного аналізу, але беззаперечним є величезне теоретичне значення цих видатних досліджень для теоретичної лінгвістики, яка збагатилася відмінковою граматикою Ч. Філлмора та дедуктивною породжувальною граматикою Н. Хомського, спонукаючи нове покоління лінгвістів до цікавих досліджень у галузі формального, функціонального і комунікативного синтаксису.

У 60–70-ті роки в епоху масового захоплення генеративізмом радянська лінгвістика пішла своїм шляхом, замінивши глобальну модель генеративної граматики моделлю «Смисл ↔ Текст» (або «Зміст ↔ Текст» І. Мельчука, О. Жолковського, Ю. Апресяна. У цій моделі природна мова розглядається як багаторівнева відповідність між смислами і текстами, при цьому надається особливе значення словнику, виявленню різних способів вираження

одного й того ж значення (лексичних, синтаксичних, логічних тощо). Розроблення семантичного компонента настільки захопило дослідників, що з'явився самостійний напрямок у семантиці, відомий як Московська семантична школа.

З початку 90-х років з'являється низка робіт Ю. Апресяна з інтегрального лінгвістичного опису та системної лексикографії, яка суттєво доповнила теорію «Смисл ↔ Текст». У цій теорії робиться акцент на взаємодії словникової та граматичної інформації, також на встановленні системних зв'язків у межах усього словника. У західній лінгвістиці ці ідеї були підтримані, і з'явилися також роботи із системного опису семантики, які вплинули на галузі, дотичні до лінгвістики – когнітивну психологію, гештальт-психологію, герменевтику, теорію прототипів. Метою цих досліджень було вироблення і формулювання загальних семантичних правил, які визначають функціонування лексем, словосполучень, речень тощо, причому правила повинні корелювати з людською поведінкою і мати антропоцентричний характер.

Теорія «Смисл ↔ Текст» та її розробники І. Мельчук, О. Жолковський, Ю. Апресян остаточно «зламали» традиційні погляди на мову, її завдання, сприяли структурному розвиткові лінгвістики та об'єднанню її з прикладною, цим самим збагатили обидві гілки структурної лінгвістики. Однак попри все не можна не погодитися з Ю. Марчуком, який справедливо зауважує, що ані теорія глибинних відмінків, ані породжувальна граматика, хоча й просувають нас до певного осмислення складу і структури речення з погляду зв'язків слів, практичних висновків для автоматизації семантики поки що не дали. Однак нові можливості комп'ютерних технологій і результати формального аналізу висловлювань, нові уявлення про штучний інтелект й інші дослідження в галузі теоретичної семантики змушують з увагою поставитися до пошуку нових спроб у досягненні практичних результатів.

Коло теоретичних та практичних завдань, так чи інакше пов'язаних з проблемами встановлення змісту тексту, надзвичайно широке й різноманітне. Результати опрацювання цієї проблеми формують третій необхідний компонент комп'ютерної граматики мови – комп'ютерну семасіологію. Розпізнавання змісту тексту становить важливу ділянку в системах так званого інформаційного пошуку, якому, у свою чергу, передує процес індексування текстів,

або їхнього розмічення за типами вміщеної в них інформації. Різновид таких систем становлять, наприклад, бібліотечні або архівні каталоги чи біобібліографічні інформаційно-довідкові системи різних установ та відомств, автоматизовані інформаційно-довідкові служби. Залежно від того, чи предмет пошуку становлять об'єкти дійсності (факти), чи описи таких об'єктів (фактів) – документи різної будови (здебільшого реферати або патенти), інформаційно-пошукові системи (ІПС) поділяють на фактографічні та документальні. Кожна інформаційно-пошукова система має спеціальну мову доступу і роботи з нею – інформаційно-пошукову мову (ІПМ). Такі ІПМ можуть становити логічну класифікацію понять тієї предметної галузі, фактів або документів якої стосується інформаційний пошук. Прикладом ІПМ класифікаційного типу, або ІПМ-класифікацій є відомі всім читачам універсальна десятична класифікація (УДК), бібліотечно-бібліографічна класифікація (ББК) або система міжнародних стандартних номерів книги (ISBN, International Standard Book Number). Спеціальні коди кожної з цих мов, які ми знаходимо на будь-якому різновиді друкарської продукції (книжці, брошурі, журналі, збірці чи газеті), і становлять інвентар її одиниць, використовуваних для індексування книжкових потоків за предметними галузями та дотичними до них поняттями. Крім універсальних ІПМ-класифікацій, є ІПМ цього типу, зорієнтовані на роботу ІПС з текстами певної предметної галузі, тематики, тобто ІПМ-класифікації спеціального призначення. Таку оригінальну ІПМ класифікаційного типу розробили автори двотомного «Словаря славянской лингвистической терминологии», виданого у Празі в 1977 р. Цей словник подає 2266 сучасних лінгвістичних термінів-понять всіма слов'янськими і трьома західноєвропейськими мовами (англійською, французькою та німецькою). Лінгвістичні терміни в ієрархічному дереві – основі цієї класифікації – розподілено за 9 предметними галузями:

I. Загальні поняття;
II. Звуковий бік мови;
III. Графічний бік мови;
IV. Словниковий склад;
V. Частини мови;
VI. Структура слова;
VII. Синтаксис;

VIII. Стиль;

IX. Нові лінгвістичні напрями і методи.

У межах кожної з цих галузей окремі терміни детально описано за додатковими, властивими їм ознаками. Подекуди така класифікаційна схема може містити 7 рівнів деталізації поняттєвої структури вихідного, базового для певної лінгвістичної галузі терміна, або 6 додаткових ознак, які уточнюють його зміст. Для позначення місця того чи іншого терміна в ієрархії понять вироблено систему спеціальних цифрових кодів. Ось, приміром, як представлені в цьому словнику-тезаурусі терміни на позначення різноманітних мовних засобів спілкування. Ці терміни перебувають в створеній укладачами словника ієрархії на 2-му рівні деталізації. Всі вони належать до галузі основних понять і тому містять цифровий код 1. Спільна для всіх них змістова ознака «засіб спілкування» в класифікації понять здобула цифровий код 5. Отже, коло всіх 23 вихідних понять цієї групи стоять тричленні цифрові коди: 1, 5 і порядковий номер терміна – назви конкретного мовного засобу спілкування. До порядкового номера, в свою чергу, можуть додаватися цифрові коди, які вказують на додаткові змістові ознаки, за якими його вміщено на відповідному місці в ієрархії лінгвістичних понять.

Перелік українських лінгвістичних термінів на позначення різновидів мовних засобів спілкування, представлених у «Словнику слов'янської лінгвістичної термінології», з відповідними цифровими кодами. природна мова 1-5-1, штучна мова 1-5-2, міжнародна мова 1-5-3, світова мова 1-5-4, інтерференція 1-5-13, 1-5-14 тощо. Максимальну (6 кроків) деталізацію лінгвістичного поняття (до 7-го рівня ієрархії) демонструють у цій групі терміни на позначення нелітературної форми мови: 1-ий рівень (код 1) мова 2-ий рівень (код 5) мовний засіб спілкування 3-ій рівень (код 23) нелітературна форма мови 4-ий рівень (код 1) діалект 5-ий рівень (код 1) наріччя 6-ий рівень (код 1) говір 7-ий рівень (код 1). Іншим способом унаочнення змісту в системах інформаційного пошуку є виділення в текстах так званих ключових слів, або слів-концептів. Такі слова в конденсованій формі виражають основну інформацію про зміст тексту. Для їх позначення використовують спеціальні одиниці – дескриптори, а тому й самі ІПМ такого типу одержали назву дескрипторних. ІПМ-класифікації та ІПМ дескрипторного типу не заперечують, а доповнюють одна одну. Мови

дескрипторного типу більше прив’язані до текстів конкретної предметної галузі або тематики, а тому виявляють більшу гнучкість та ефективність у процесі автоматичного аналізу їхнього змісту.

В ІПМ дескриптори можуть становити окремі слова, словосполуки або й частини слів, які виражають засадничі для окремої предметної галузі поняття. Для упорядкування інвентаря дескрипторів, а також уніфікації позначення понять у кожній ІПС створюється спеціальний інформаційно-пошуковий тезаурус (ІПТ), який складає лінгвістичне забезпечення (lingware) такої системи. Дескриптори в такому ІПТ упорядковують на основі не лише парадигматичних, а й синтагматичних відношень. Саме завдяки урахуванню останніх в ІПТ увиразнюють відношення так званої квазісинонімії, або контекстної синонімії, коли дескриптори на позначення певних понять зближуються лише в текстах, що стосуються окремої предметної галузі або певної проблемної ситуації в її межах. Крім того, до ІПТ потрапляють і так звані асоціативні дескриптори, тобто слова, що можуть виявляти лише опосередковану семантичну близькість у певних комунікативних ситуаціях. Скажімо, дескриптор детство в «Російському семантичному словнику» за редакцією Ю. М. Караулова перебуває в опосередкованих, асоціативних зв’язках з дескрипторами «наив» та «нерозум», що позначають поняття «наивность» та «неразумность», пор. такі визначення, як дитячий погляд на речі (=наївний) та поводитися по-дитячому, як дитина (=нерозумно, нерозсудливо). Процедура пошуку інформації в ІПС здійснюється в режимі «запит – відповідь». «Запит» на пошук інформації містить спеціальний пошуковий образ документа (ПОД), який створюють вручну або за допомогою комп’ютера, індексуючи (розмічаючи) текст з допомогою одиниць певної ІПМ, зокрема дескрипторів. «Відповідь», або пошуковий припис (ПП) на такий «запит» становить певним чином упорядкована сукупність дескрипторів, які описують певну проблемну ситуацію або предметну галузь у цілому в ІПТ системи. Після порівняння ПОД та ПП користувач ІПС одержує всі документи певної бібліотеки, архіву або взагалі будь-якого інформаційного масиву, зміст яких відповідає вміщеним у ПОД та ПП дескрипторам або одиницям (наприклад, кодам) мов-класифікацій. При цьому основними вимогами до ПОД та ПП є повнота та точність видачі інформації. Чим вищі параметри повноти й точності інформаційного пошуку, тим менший у такій

системі показник інформаційного шуму, або неправильно виданої у відповідь на запит інформації. Для усунення інформаційного шуму застосовують методики індексування тексту, які враховують комунікативну значущість та функціональне навантаження слів у ньому. Одну з таких методик виділення в тексті ключових слів на основі процедур сіткового моделювання лексики розробив український дослідник Е. Ф. Скороходько.

В семантичній сітці слова впорядковуються залежно від того, які вони мають семантичні складники (компоненти) або дериватами якого іншого слова вони виступають. Отже, найбільше функціональне навантаження в тексті матимуть слова, що містять найбільшу кількість семантичних складників або з них можна вивести найбільшу кількість семантичних дериватів, Таким словам у тексті під час індексування приписують найбільшу вагу, або ранг. Наприклад, слово-родова назва, або гіперонім, одержить більшу вагу (вищий ранг), ніж слово-видова назва (гіпонім).

Найефективнішими для потреб індексування тексту виявилися гнучкі методики встановлення ключових слів, які поєднують різні функціональні властивості слів: їхню частоту (абсолютну, середню й відносну), комунікативну значущість, силу зв’язків з іншими словами в тексті (словотвірних, синтаксичних, асоціативних) тощо.

Цікавий підхід до індексування лінгвістичних текстів реалізували в своїй праці польські дослідниці З. Руднік-Карватова та Х. Карпінська. Проіндексувавши тексти авторефератів мовознавчих дисертацій, вони уклали «Словник ключових слів славістичного мовознавства». У цьому словнику близько 2500 термінів з різних галузей сучасної славістики впорядковано за абеткою, між ними встановлено парадигматичні відношення (переважно родо-видові та відношення за ознакою «частина-ціле») та синтагматичні (синонімічні) відношення. Наприклад, для родового терміна – ключового слова лінгвістичних текстів мова (пол. język) укладачі словника встановили 90 видових термінів – ключових слів текстів, присвячених окремим лінгвістичним проблемам. Серед таких видових ключових слів назви різновидів мови за походженням або належністю до певної групи чи родини, за генеалогічною класифікацією мов, за типом будови мови, за сферою суспільного життя, яку обслуговує мова тощо. Синонімічні відношення, або відношення рівноправності в ієрархії ключових слів виявили такі пари лінгвістичних термінів, як мова засобів

масової інформації – мова медіа, мова масмедіа або мова етнічна – етнолект. Дібрані польськими дослідницями ключові слова досить детально структурують інформаційне поле сучасної лінгвістики. Спирання на них в процесі пошуку повинно забезпечити високий ступінь точності одержуваної інформації. Великого поширення набули різноманітні методики так званого контент-аналізу, або аналізу змісту тексту за певними концептуальними змінними, що позначають центральне поняття аналізованого тексту. Активно такі методики логіко-семантичного аналізу тексту застосовують останнім часом в дослідженнях із політичної лінгвістики, піартехнологій.

Саме нові ефективні методики встановлення у тексті ключових слів дали можливість застосовувати в системах ІПС метод так званого безтезаурусного пошуку. Він передбачає роботу з масивами документів в інтерактивному режимі з допомогою спеціальної діалогової системи, що дозволяє користувачеві створювати потрібні пошукові образи документів під час безпосереднього перегляду того чи іншого інформаційного масиву й залежно від типу опрацьовуваних документів вносити в такі пошукові образи необхідні корективи.

Питання для самоконтролю

1. Які основні проблеми автоматичного семантичного аналізу?
2. Назвіть два напрямки формалізації семантики.
3. Які основні принципи автоматичного семантичного аналізу?
4. В чому специфіка категоризації лексики та семантичні характеристики слів, що застосовуються при автоматичному семантичному аналізу мови?
5. Що таке семантична сітка?
6. Що таке модель «м’якого» розпізнавання тексту комп’ютером?
7. Проаналізуйте лінгвістичний та інформаційний підходи до семантичного представлення тексту.
8. Що таке семантичний компонент аналізу тексту?
9. Який склад семантичного компонента?
10. Що таке метамова семантичних структур?

Завдання

1. Зробіть мультимедійний огляд програм семантичного аналізу та елементів автоматичної обробки семантики в різних прикладних програмах.
2. Порівняйте специфіку роботи сервісів, визначте їх основні переваги та недоліки:

– http://www.analyst.ru/index.php?lang=eng&dir=content/products/&id=ta)
– http://www.galaktika-zoom.ru
– http://download.cnet.com/NetXtract-Personal/3000-12512_4-10073214.html

ТЕМА 8
АВТОМАТИЧНИЙ МОРФОЛОГІЧНИЙ АНАЛІЗ

План

1. Лінгвістичні засади створення автоматичного морфологічного аналізу.
2. Експериментальні та промислові системи автоматичного морфологічного аналізу.
3. Доморфологічний аналіз (виділення в тексті одиниць, які не мають морфологічного статусу).
4. Типологія алгоритмів автоматичного морфологічного аналізу.

Теоретична частина

Обов’язковою складовою частиною лінгвістичного забезпечення будь-якої системи автоматичного опрацювання тексту є автоматичний морфологічний аналіз (АМА), до завдань якого входять: визначення для одиниць тексту інформації про місце їх у морфологічній системі відповідної мови; ідентифікація словоформ однієї лексеми.

Внаслідок роботи АМА кожній словоформі тексту приписуються коди частин мови і значення граматичних категорій (рід, число, відмінок, вид, час, особа тощо). Характер цієї інформації, обсяг її й методи, за допомогою яких встановлюється морфологічна інформація, залежать від мети дослідження, у межах якого здійснюється АМА, від орієнтації на характер текстів, що аналізуються. Морфологічний аналіз присутній на всіх етапах аналізу тексту, тому що ані морфемний, ані синтаксичний, ані

семантичний аналіз не можуть обійтися без визначення частин мови. Наприклад, при морфемному сегментуванні тексту виділення префіксів можливе без знання частин мови, а суфіксів – ні: потрібні різні їх набори та процедури відсікання суфіксів для іменників, дієслів, прикметників, прислівників. При автоматичному синтаксичному аналізі лише за наявності лексико-граматичної та граматичної інформації до кожної словоформи можна синтаксично прив'язати словоформи у реченні. На рівні формального аналізу тексту саме морфологічна інформація забезпечує доступ комп'ютера до змісту, опосередкованого через співвіднесеність одиниць змісту з одиницями виразу. Наприклад, у реченні *Студенти уважно слухають викладача* визначення комп'ютером морфологічної частиномовної належності через певні формальні граматичні ознаки (*студенти, викладача* – іменники, *слухають* – дієслово, *уважно* – прислівник) одержуємо первинну семантичну характеристику, оскільки іменник узагальнено позначає *предмет,* дієслово – *дію*; прикметник – *ознаку;* прислівник – *ознаку ознаки.* Наступне уточнення категоріальних ознак називного відмінка множини у *студенти,* знахідного однини у *викладача,* часово-особових – у *слухають* тощо наближують нас до розуміння синтаксичної будови речення: підмет (*студенти,* тому що наз. відм.), присудок *(слухають,* тому що особова форма), додаток *(викладача,* тому що знах. відм.), обставина *(уважно),* а звідси інформація про суб'єкт, об'єкт дії, детермінанти. Зрозуміло, що до здобуття смислу із цього речення ще довгий шлях, але починається він із морфологічної частиномовної інформації.

Морфологічні ознаки одиниць тексту мають стати інструментом дослідження зв'язку між лексикою і граматикою, між використанням його у мовленні, між парадигматикою (в аспекті розгляду відмінкових форм відмінюваних слів) і синтагматикою (в аспекті лінійних зв'язків слів, сполучуваності у тексті). Роль саме такого «перекидного містка» виконують частини мови.

Звернемося до характеристики основних понять морфології як розділу граматики будь-якої мови.

Морфологія – це одна із частин граматичної будови мови, що охоплює граматичні класи слів (частини мови), граматичні (морфологічні) категорії цих частин мови та їхні форми.

Морфологія як наука передбачає розв'язання таких завдань:

1) вивчення граматичних класів слів – частин мови і

принципів їхнього класифікаційного виділення;

2) виокремлення частини семантики слова як морфологічної;

3) обґрунтування набору морфологічних категорій та їхньої природи;

4) опис сукупності формальних засобів, закріплених за відповідними частинами мови та їхніми морфологічними категоріями.

Принципову відмінність стратегій традиційного і комп'ютерного морфологічного аналізу визначив Ю. Марчук: у комп'ютерній лінгвістиці поняття морфологічного аналізу є поняттям операційним. Якщо у традиційній лінгвістиці до морфологічного аналізу належить те, що характеризує форму і відповідає на питання «що» класифікують, то в обчислювальній (прикладній) лінгвістиці важливо не «що», а «як» одержують ту чи іншу інформацію. Справді, до етапу морфологічного аналізу входить велика кількість операцій, за допомогою яких можна одержати необхідну морфологічну інформацію.

Практично АМА присутній в усіх видах аналізу тексту, оскільки жоден із них не може обійтися без аналізу форм слів, визначення належності слова до граматичного класу. Лінгвістичним поясненням цього може бути об'єктивно існуючий тісний зв'язок між лексичними і граматичними значеннями одиниць мови, а також між системами парадигматичних і синтагматичних відношень. Як зазначали Н. Гендіна і Д. Хейс, саме морфологічна інформація забезпечує доступ комп'ютеру до змісту тексту, оскільки досі єдиним реальним шляхом автоматичного аналізу плану змісту залишається опосередкований шлях через співвіднесення його з одиницями плану вираження. На думку А. Савченко, погляд на частини мови як на лексико-граматичні класи слів слугує теоретичною підставою обов'язковості розв'язання задачі визначення граматичних класів при автоматичній переробці текстової інформації.

Розв'язання проблеми автоматичного кодування слів тексту, тобто приписування їм кодів граматичних класів, пов'язане, насамперед, із питанням принципів граматичної класифікації. Справді, перш ніж виділяти, треба знати, що виділяти, а потім – як виділяти, тобто знайти такі формальні ознаки, з якими може працювати комп'ютер.

Основною одиницею морфології, як і лексикології, є слово,

але предметом вивчення лексикології є лексичне значення – предметно-речовий зміст слова, а предметом вивчення морфології є граматичне значення – показник різних відношень, у які вступає слово з іншими словами у словосполученні, реченні, тексті. Об'єктом морфології є структура слова, форми словозміни, способи вираження граматичних значень.

Словоформи – це граматичні форми одного й того ж слова, тотожні лексично (мають спільне лексичне значення), але протиставлені граматичним значенням: *пишу, пишеш, пише, пишемо, пишете, пишуть, писав, писала, писало, писали, пиши, пишімо, пишіть, писатиму, буду писати.* Ці звукові комплекси мають спільне лексичне значення, це словоформи, які є конкретними представниками слова у мовленні. Слова змінюють форму відповідно до закономірностей сполучуваності в реченнях, залишаючись при цьому в межах тих самих лексичних значень. *Морфологічне слово* – де сукупність граматичних форм слів, або словоформ. Упорядкована сукупність граматичних форм слова називається парадигмою**.**

Граматичні значення органічно входять до семантичного плану кожного окремо взятого слова, обов'язково супроводжують і підтримують його вираження. Так, слово *студент* – це не тільки послідовність із семи звуків, якою в українській мові передається назва особи, яка навчається у вищому або середньому спеціальному навчальному закладі, а і єдність трьох граматичних (морфологічних) ознак – значень роду (чоловічий рід), числа (однина) і відмінка (називний). У формі слова *викладача* поєднуються граматичні значення чоловічого роду, однини, знахідного відмінка.

Розгляньмо, з яких елементів складається граматична (морфологічна) інформація, зосереджена в дієслові-присудку *слухають.* Із цієї форми випливає, що відповідна процесуальна ознака реалізується одночасно з моментом мовлення про неї, не є завершеним, доведеним до певної межі процесом, не пов'язана ні з тим, хто говорить, ні з адресатом мовлення *(ти),* вказує лише на понад один предмет *(студент-и*), що виконують певну дію. Наведеним ознакам відповідають граматичні значення теперішнього часу, недоконаного виду, третьої особи множини.

Граматичні значення набагато ширші, абстрактніші порівняно із власне лексичними, які мають індивідуальний характер. Такі далекі із семантичного погляду іменники, *як тінь, відстань,*

антресоль, модель, акварель, заповідь, повінь, піч, ніч, деталь, верф тощо, нічим не відрізняються щодо граматичних значень від іменника *осінь* – усі вони іменники жіночого роду, однини, називного відмінка. За таким же принципом узагальнення виділяються і дієслівні граматичні значення: дієслова *летіти* і *сидіти* позначають різні, навіть різко протилежні процеси, але вони абсолютно тотожні, рівнозначні як між собою, так і з дієсловом *слухати* з погляду граматичних характеристик.

Таким чином, можна констатувати: під граматичним (морфологічним) треба розуміти значення, абстраговане внаслідок обов'язкового розрізнення не менше двох однотипних, постійно повторюваних ознак великої кількості конкретних слів із властивими їм лексичними значеннями.

Спеціальні постійно використовувані засоби мовного вираження граматичних значень називаються граматичними формами. Ці два поняття перебувають у найтіснішій єдності: граматичне значення знаходить свій вияв тільки в тій або іншій формі, яка, у свою чергу, покликана передавати граматичне значення. Суттєвою особливістю цієї єдності є те, що одне й те ж граматичне значення може мати різні матеріальні засоби вираження, або форми. Наприклад, в українській мові співіснують дві рівноцінні з граматичного погляду варіантні форми майбутнього часу *(читатиму* і *буду читати),* дві форми давального відмінка (*вчителю, вчителеві).*

АМА призначено для писемного тексту. Різні мови користуються різними системами письма (буквеними, складовими тощо). Крім того, важливими є відомості про те, як співвідносяться усне і писемне мовлення (напр., у писемному тексті відсутні деякі голосні або присутній наголос, який має морфологічне значення тощо). Істотним є також спосіб членування тексту спеціальними засобами (пробіл, пунктуаційні знаки, величина букв, знаки членування тексту на структурні частини тощо).

Тематика тексту. Кожному тексту як результату мовленнєвої діяльності й засобу комунікації відповідає певна система понять, що відображають його тематичну спрямованість. Дослідження лексичного складу, морфологічних характеристик, синтаксичних структур текстів різної тематичної спрямованості виявляють розбіжності у вживанні лінгвістичних одиниць, що іноді враховується при автоматизації морфологічного аналізу.

Системи АМА створювалися спочатку як *експериментальні.* Лінгвісти перевіряли правильність і достатність прийнятих лінгвістичних теорій з метою автоматизації. У перших системах АМА, які розроблялися у 50-60-х рр., використовувалися дані з теорії морфемного аналізу, морфонології, парадигматики лексико-граматичних класів слів. Автоматичний аналіз у таких системах розпадається на три складові частини: автоматичне виділення основи у словоформі тексту; пошук основи у словнику основ; порівняння структури словоформи з даними про її основу, які містяться у словнику основ.

Іншими словами, кожна словоформа тексту аналізується за допомогою заздалегідь укладених словників основ, коренів, префіксів, суфіксів, флексій. Омонімія словоформ не розрізнялася.

Під час роботи над експериментальними системами АМА із часом вироблялися нові принципи роботи з комп'ютером: у системах АМА з'явилася можливість спростити процедуру визначення морфологічних характеристик словоформ з урахуванням «інтелектуальних» можливостей комп'ютера, виявилася перспективною ідея використання текстових закономірностей зживання словоформи, поєднання морфологічного аналізу із синтаксичним, а в деяких випадках із семантичним.

Процедура автоматичного індексування розбита на блоки, які працюють послідовно: морфологічний аналіз, синтаксичний аналіз, семантико-синтаксичний аналіз прийменникових конструкцій та варіювання смислового запису запиту.

У 70-80-ті рр. XX ст. почали створюватися так звані *промислові* системи опрацювання текстової інформації. Як зауважує Ю. Марчук, це системи автоматичного перекладу з однієї мови іншою, системи інформаційного пошуку літератури, автоматизоване й автоматичне редагування текстів, автоматизоване анотування і реферування літератури.

Однією з розробок засобів індексування тексту документів є система SMART, описана Г. Селтоном. Основу системи складає розвинена система словників і засоби їхнього обслуговування.

Метод індексування на основі семантичного аналізу розглядається у роботі Н. Леонтьєвої та С. Вишнякової. Процедура індексування складається із двох етапів: індексування за дескрипторним словником (режим «слово»); індексування за допомогою автоматичного інформаційно-пошукового тезауруса.

Дескрипторний словник має структуру таблиці, що складається із трьох колонок. У першій записуються основи слів; у другій – набори дескрипторів, приписані кожній основі (під дескриптором автори розуміють елементарне поняття: кожне слово розуміється як набір цих дескрипторів); у третій – граматичні ознаки дескрипторів. У режимі «слово» переведення тексту реферату на інформаційну мову відбувається із частковим морфологічним аналізом та лематизацією.

В Україні над створенням автоматичного морфологічного аналізу російського тексту із середини 80-х рр. працював колектив співробітників відділу структурно-математичної лінгвістики Інституту мовознавства ім. О.О. Потебні НАНУ під керівництвом д-ра філол. наук В. Перебийніс у складі канд. філол. наук Т. Грязнухіної, канд. філол. наук Н. Дарчук, канд. філол. наук А. Комарової, канд. філол. наук В. Критської, канд. філол. наук А. Орлової, Л. Братищенко, Т. Пуздирєвої, теоретичні засади якого детально описані в монографії «Морфологический анализ научного текста на ЭВМ».

У багатьох системах обробки тексту існує етап, який прийнято називати доморфологічним. Для виділення одиниць аналізу треба визначитися, що таке слово? Чи будуть пунктуаційні знаки або аналітична форма майбутнього часу *буду читати* словом? Яке значення великої літери на початку речення, у власної назви, в абревіатурі тощо? Яке функціональне значення крапки: це або кінець речення, або позиція біля скорочення, або біля рубрикації тощо.

Машинними словами вважаються ланцюжки графем від пробілу до пробілу, у тому числі пунктуаційні знаки.

Крапка репрезентує кілька ситуацій: скорочені слова, які є типовими в текстах, вони задані у словнику; рубрикацію (після цифри або букви), що враховується правилами алгоритму доморфологічного аналізу; ініціали, що враховуються також правилами алгоритму (відсутність пробілу між скороченням імені, по-батькові, які знаходяться у препозиції або постпозиції до прізвища: *В. Іванов або Іванов В.І.)*.

У решті випадків крапка вважається кінцем речення й отримує відповідну позначку.

Слово, яке стоїть після крапки і пробілу, вважається початком речення і перед ним на етапі доморфологічного аналізу ставиться

позначка початку речення.

У назвах тексту, навпаки, не ставиться крапка, але для багатьох задач (напр., автоматичне реферування) ця інформація потрібна, тому в кінці назви ставиться умовна крапка і фіксується умовний кінець речення (якщо потрібна інформація про назву, вона також фіксується).

Дробові числа мають написання: з комою і з крапкою, що також ураховується алгоритмічними правилами.

Усі слова, що мають у своєму складі одну або кілька великих букв (російські, якщо аналізується російський текст, чи українські, якщо аналізується лише український текст) у сполученні зі знаками дефіс (*міні-ЕОМ),* тире *(М – числення*), або скісну риску *(введення/виведення)* (іноді помилково дефіс вживається замість тире і навпаки), розглядаються окремо, тобто аналізуються як окремі лексичні одиниці. Те ж стосується композитів, наприклад, слово *семантико-синтаксичний* алгоритмічно членується на *семантико-* і *синтаксичний, вагон-ресторан,* які представлені як окремі одиниці словника з відповідними характеристиками до нього (для *семантико-* немає жодної характеристики, тому що воно є частиною композита, а *синтаксичний* отримує код ад'єктива; *вагон* і *ресторан* одержують відповідно коди іменника чол. р.), причому дефіс між *вагон* і *ресторан* алгоритмічно запам'ятовується, щоб на наступних етапах опрацювання тексту (машинний переклад) його відновити.

Слова, що починаються ланцюжком цифр або латинських букв, за якими йде дефіс, а післядефісна частина складається з курсивних літер, властивих аналізованому тексту (російські або українські), також членуються і післядефісна частина для ідентифікації потрапляє до словника *(п-розрядний, 2-го).*

Слово, яке складається з букв латинського алфавіту, отримує окремий код (напр., ФФ). Якщо у слові російського тексту вживаються українські літери (можливо, це українське слово або помилка у слові) і навпаки, в українському тексті – відсутні в його алфавіті літери, то слову присвоюється окремий код-знак (?).

Зрозуміло, що укладання алгоритму доморфологічного аналізу досить копітка, але необхідна робота, цей етап присутній у багатьох відомих нам системах АМА, у той же час сам алгоритм може бути прийнятий за основу коректора тексту, який виявляє помилки на етапі доморфологічного аналізу.

Правильність і повнота результатів аналізу тексту в системах АОТ залежить від кількох факторів: від рівня знань про мову і мовлення, тобто правильності лінгвістичної теорії, покладеної в основу АМА; від рівня формалізації цих знань у створюваній комп'ютерній граматиці.

У сучасних системах АМА існують два основних принципи виведення морфологічних ознак слова за допомогою його структури:

1) здобуття граматичної інформації зі слова шляхом його графемного аналізу;

2) представлення граматичної інформації у словнику основ і словнику флексій.

Перший принцип базується на граматичних ознаках, які містяться у кінцевих буквосполученнях, а другий – на граматичній інформації у словнику основ і словнику флексій, моделюючи класичну схему аналізу шляхом поділу словоформи на основу і передбачувану флексію з наступною перевіркою на сумісність флексії з основою.

У перші роки роботи зі створення машинного перекладу було запропоновано велику кількість різних алгоритмів АМА для різних мов з різними морфологічними системами.

Алгоритм виділення іменників та прикметників з тексту представлено у роботі Н. Кравченко. У статті наводиться список усіх неіменникових кінцевих буквосполучень, які умовно називаються закінченнями, і всіх іменникових кінцевих закінчень, останні букви яких збігаються з якимось неіменниковим закінченням. Цей список містить 130 закінчень. Використовується ще й додатковий список із 500 словоформ, які не мають формальних ознак приналежності до граматичного класу.

У списку-фільтрі закінчень алгоритму В. Отрадинського нараховується 120 закінчень.

Метод графемного аналізу в Україні було застосовано для розробки АМА російськомовних науково-реферативних текстів і текстів української мови. Вибір текстів (російськомовні науково-реферативні з кібернетики й обчислювальної математики та українські наукові тексти) не був випадковим. З одного боку, орієнтація на реферативні тексти пояснювалася важливою роллю вторинних документів у процесі оперативного обміну інформацією, що посилювало актуальність і практичну значимість системи у

плані використання її при опрацюванні інформації. З іншого, як показує досвід працюючих систем АОТ, у системах, орієнтованих на оброблення науково-технічних текстів широкої тематики, які охоплюють різні підмови, АМА спирається на словникову інформацію, а для систем, орієнтованих на конкретні підмови або працюючих із текстами вторинних документів (рефератів, патентів тощо), які відрізняються обмеженою лексикою і характеризуються стандартизованістю морфології (обмеженість у вживанні граматичних форм, синтаксичних структур), доцільною була методика графемного аналізу.

У мовах флективного типу, з розгалуженою системою словозміни, до яких належить російська й українська, інформація про граматичні значення зосереджена в кінці слова і формально виражена флексією чи формотворчим суфіксом, тому аналіз здобув назву флективного аналізу (ФА). Основним інструментом АМА як засобу ідентифікації граматичної інформації є список КФА – кінцівок словоформ, що дозволяли однозначно встановлювати частиномовну належність словоформ тексту та їх граматичну характеристику.

У системі АМА розроблено дві версії укладання списку КФА. Перша передбачає автоматичне формування списку за вибіркою текстів, закодованих попередньо вручну в термінах граматичних класів. При цьому передбачається багаторазове автоматичне коригування списку за словоформами, доданими до вихідної вибірки, які були не розпізнані або розпізнані неправильно на незакодованому тексті. Цей принцип називається принципом навчальної вибірки, її обсяг і кількість переформувань списку встановлюються емпіричним шляхом.

Друга версія (прийнята як робоча) передбачає формування списку КФЛ вручну на основі лінгвістичного аналізу оберненого словника словоформ, який укладається автоматично на певній вибірці текстів, з урахуванням даних Граматичного словника російської мови для АМА російської мови або Оберненого словника для української.

Правила алгоритму, за якими вихідна словоформа тексту перевіряється за списком, побудовані таким чином, що у списку допускається вкладання КФЛ один в один, оскільки порівняння починається з найдовшої «кінцівки» при збігу останньої графеми у текстової словоформи й у КФЛ зі списку.

У розробці морфологічного аналізу виділилося кілька напрямів. Один із них моделює класичну схему аналізу шляхом поділу словоформи на основу і ймовірне закінчення з подальшою перевіркою на сумісність закінчення з основою, яка залишається. Інший напрям використовує інформацію, що міститься в кінцевих буквосполученнях. Ця інформація утворюється в результаті попередньої статистичної обробки словника. Третій напрям створює універсальні математичні моделі морфології у формі відкритих систем рівнянь, що дозволяють шляхом обчислення здійснювати нормалізацію словоформ, отримання граматичної інформації і синтез словоформ.

В основу побудови алгоритмів морфологічного аналізу покладено розбиття всіх слів на класи, що визначають характер зміни літерного складу форм слова. Ці класи можуть бути названі морфологічними. Зміни форм слів можуть носити різний характер. Вони можуть бути пов’язані як з зміною основи слова, так і зі змінами його закінчення .

Морфологічні класи слів діляться на два види: основозмінні класи, що характеризують систему зміни основ; флективні класи слів (класи незмінних слів виділялися тільки за синтаксичним принципом).

За своєю синтаксичної функції змінювані слова об’єднані в такі групи: іменники; прикметники; дієслова в особовій формі; дієслова минулого часу, короткі прикметники і дієприкметник; кількісні числівники.

Флективний клас може бути охарактеризований або певною системою ознак, або словом-представником, яке є носієм цих ознак. Ознаками, за якими змінюване слово може бути віднесено до певного класу, є: належність до однієї із синтаксичних груп (або підгруп); система закінчень (тип словозміни).

Усі наявні в галузі машинної морфології розробки можна розділити на дві групи залежно від принципової орієнтації: використання словника словоформ або словника основ у сполученні з словником флексій; відмова від словників.

Однією з ґрунтовних робіт був алгоритм визначення граматичних класів слів, описаний у статті Г. Бєлоногова й І. Давидової. У ній наводяться списки кінцевих буквосполучень, що містять від однієї до п’яти графем.

Алгоритм виділення іменників та прикметників із тексту подано в роботі Н. Кравченко.У статті наводиться список усіх неіменникових кінцевих буквосполучень, які умовно називаються закінченнями, і всіх іменникових кінцевих закінчень, останні букви яких збігаються з якимось неіменниковим закінченням. Також використовується ще й додатковий список словоформ, які не мають формальних ознак належності до граматичного класу.

Досвід роботи над графемним аналізом російського тексту було узагальнено у книзі Г. Бєлоногова і В. Богатирьової «Автоматизовані інформаційні системи», де подано таблиці для визначення граматичних класів за кінцевими буквосполученнями словоформ. Для достовірного однозначного визначення граматичного класу алгоритмічно використовуються дані про останні 2-4 букви словоформ.

Основним інструментом автоматичного морфологічного аналізу, як засобу ідентифікації граматичної інформації є список квазіфлексій – кінцівок словоформ, що дозволяли однозначно встановлювати частиномовну належність словоформ тексту та їхню граматичну характеристику.

Розроблено дві версії укладання списку квазіфлексій: принцип навчальної вибірки. Передбачає автоматичне формування списку за вибіркою текстів, закодованих попередньо вручну в термінах граматичних класів; вибірка з урахуванням даних словників. Передбачає формування списку вручну на основі лінгвістичного аналізу зворотного словника словоформ.

Правила алгоритму, за якими вихідна словоформа тексту перевіряється за списком, побудовані так, що у списку допускається вкладання квазіфлексій один в один, оскільки порівняння починається з найдовшої «кінцівки» за збігу останньої графеми у текстовій словоформі й у квазіфлексіях зі списку.

Попередні дослідження взаємозв’язку між графемною структурою словоформи і такими її морфологічними характеристиками, як рід, число, відмінок іменних класів, час, особа, число для дієслова, показали, що аналіз графемної структури словоформи може бути використаний не лише як інструмент ідентифікації лексико-граматичних класів у тексті, а і у визначенні граматичних підкласів (у межах класу). В описуваній системі автоматичного морфологічного аналізу такий аналіз реалізується на тих самих принципах флективного аналізу за допомогою списків

квазіфлексій, які використовуються на етапі визначення класів слів. У його завдання входить опрацювання текстових одиниць (узятих окремо, поза контекстом) для визначення для кожної з них набору можливих граматичних значень, які вияскравлюють словозміну й узгодження з іншими одиницями в тексті і виявляють підклас певного коду лексико-граматичного класу слів.

Підклас є відбиттям морфологічних і частково певних синтаксичних характеристик одиниць граматичного класу, наприклад, клас прийменників є ефективним засобом організації синтагматичних зв'язків слів, тому він має підкласи, що відбивають лише синтаксичні властивості.

Формування списку квазіфлексій передбачає розв’язання питання про можливість однозначного визначення морфологічної інформації на основі графемної структури слова. Усе ж, по-перше, жодний аналіз видобутої з контексту словоформи не може в усіх випадках визначити однозначно належність її до певного лексико-граматичного класу слів, і, по-друге, немає жодної словозмінної парадигми слова іменних лексико-граматичних класів слів, у якій була відсутня омонімія словозмінних форм. Ці дві обставини змусили ввести поняття диз'юнктивних кодів класів і підкласів, які будуть аналізуватися на наступному етапі контекстного аналізу.

Підклас визначається для тих одиниць тексту, які у визначенні граматичного класу одержали однозначний (недиз’юнктивний) граматичний, а також для словоформ з диз'юнктивними кодами, що вияскравлюють омонімію класів з однаковими словозмінними характеристиками.

Підкласи різних граматичних класів відрізняються за своєю граматичною природою. У змінюваних класів слів вони є виразниками суто морфологічних характеристик. У службових частин мови – прийменника і сполучника – підклас вказує на їхні синтаксичні особливості, а саме: код підкласу прийменника містить інформацію про відмінки, якими цей прийменник може керувати; код підкласу сполучника вказує на тип зв’язку, що формується за участі сполучника. Обидва типи інформації необхідні для синтаксичного аналізу і враховуватимуться на етапі контекстного аналізу для зняття омонімії повнозначних частин мови. «Розпізнавальна сила» квазіфлексій, за якими здійснюється ідентифікація граматичних підкласів, у різних частин мови різна.

Для визначення однозначної інформації для омоформ потрібні додаткові дані з тексту, які допоможуть зняти омонімію роду, числа, відмінка.

Розрізняють такі види морфологічного аналізу зі словником: морфологічний аналіз зі словником основ; морфологічний аналіз зі словником словоформ; морфологічний аналіз методом логічного множення; морфологічний аналіз без словника, за допомогою таблиць.

Автоматичний морфологічний аналіз зі словником основ. У цьому виді аналізу використовується словник основ слів і ряд допоміжних таблиць. У словник включені основи простих і складних слів без внутрішньої флексії. Якщо слово має кілька форм основ, то в словник включені всі форми основ слів. Кожній основі словника ставиться у відповідність поєднання коду основозмінного класу й коду флективного класу, а омонімічній основі – серія поєднань таких кодів. Так влаштований словник у системі, описаної Г. Г. Бєлоногови м.

Морфологічний аналіз слова починається з його флективного аналізу. Останній проводиться для правильного виділення основи, заміни літерного складу порядковим номером за словником і визначення граматичної інформації слова. Алгоритм морфологічного аналізу складається з 32 блоків і враховує всі кроки морфологічного аналізу за допомогою словника основ, можливі варіанти аналізу при відхиленні процесу від однозначних правил, перехід до наступних ступенів аналізу.

Морфологічний аналіз зі словником словоформ застосовується, коли морфологія певної мови досить бідна. Крім того, на перший погляд видається, що алгоритм аналізу зі словником словоформ простіший, ніж алгоритм роботи зі словником основ: не треба здійснювати членування вхідний словоформи на морфеми з послідовним пошуком за словником тощо. Однак аналізі зі словником словоформ має такі проблеми: аналіз не знайдених у словнику словоформ, адже визначення інформації для словоформи, не знайденої у словнику, є необхідним для наступного етапу аналізу, коли треба принаймні визначити частину мови; ототожнення різних словоформ того самого слова:якщо кожна словоформа буде виступати як самостійна лексична одиниця, то це істотно ускладнить весь наступний аналіз і синтез. Словоформи одного слова повинні бути виокремленні як

такі. Це означає, що система морфологічного аналізу зі словником словоформ повинна мати список афіксів, коренів (основ) слів та інші необхідні атрибути для ідентифікації різних словоформ однієї і тієї ж лексичної одиниці.

Ці вимоги фактично зводять нанівець переваги аналізу зі словником словоформ, і тому аналіз зі словником основ застосовується значно частіше.

Морфологічний аналіз методом логічних множень. Завдання морфологічного аналізу вимагає суворого математичного формулювання. С. Я. Фітіаловим були викладені загальні положення побудови формальної морфології. Нехай задано словник словоформ, що описує кожну словоформу множиною -sy вигляді якоїсь інформаційної функції F(S).

Потрібно створити: множину морфем, тобто зв'язаних непустих частин словоформ, на які розчленовується будь-яка словоформа з вихідного словника; алгоритм розчленовування кожної словоформи на її складові – морфеми; словник морфем, що містить інформацію опису F(S); спосіб одержання інформації про словоформу з інформації про морфеми даної словоформи.

На базі цих словникових словоформ створюється словникова функція. Розв'язання цієї функції має вигляд об'єднання інформацій про всі словоформи. Такого роду об'єднання інформацій відповідають логічній функції диз'юнкції.

Сутність методу зводиться до таких положень. Спочатку провадиться пошук слова в словнику основ. Якщо слова, які мають закінчення, не знаходяться в словнику, тоді від кожного такого слова відкидається одна буква справа і пошук триває. При негативній відповіді відкидається наступна буква і т.д. Кожна відкинута буква вважається закінченням і фіксується як одиниця морфологічного аналізу. Їй приписується нульовий вектор – як сукупність нулів і одиниць.

Кількість компонентів векторів дорівнює кількості граматичних категорій, що можуть бути виражені закінченням, суфіксом тощо. Оскільки попередньо був зроблений пошук у словнику основ і було встановлено, якою частиною мови є аналізоване слово, маємо можливість класифікувати їх за флексіями.

Наприклад, потрібно визначити, в якому числі і відмінку стоїть слово «столом». Після пошуку в словнику було встановлено,

що «стіл» – основа іменника, а букви, що належать до закінчення, -о та -м. Буква «м» зустрічається серед букв закінчень іменника в орудному відмінку однини, чоловічого і середнього роду, а також у давальному й орудному відмінках множини всіх трьох родів. Букві «о» приписуємо такий вектор, у якому одиниці стоять у тих відмінках, у яких вона зустрічається (орудний відмінок однини чоловічого і середнього роду, орудний відмінок жіночого роду), і нулі там, де букви «о» немає.

Здійснивши логічне множення векторів, одержимо в підсумковому векторі одиницю на місці розряду тієї граматичної категорії, у закінченні якої зустрічаються паралельно і буква «о», і буква «м», а саме: у розряді орудного відмінка однини чоловічого роду.

Незалежний морфологічний аналіз. Під незалежним аналізом розуміють вивчення комбінаторики флексій і афіксів поза основами і відповідними словами. В алгоритмі немає спеціального словника основ. Ідея полягає в максимальному використанні інформації флексій у флективних мовах уже на першому етапі аналізу. Група флексій характеризується однаковим набором граматичних відношень, які передаються в суть утворення морфеми. Флексії, що входять в одну морфему, є аломорфами – (-а, -я – для українських прикметників). Завдання алгоритму полягає в тому, щоб за взаємним розташуванням аломорфів у фразі віднести кожну флексію до її морфеми. Для вирішення цього завдання укладають спеціальні словники: слів, що не несуть ніякої граматичної інформації (прислівники); флексій, де кожна графема має вказівку на те, у які морфеми вона входить; словник службових слів (прийменники та ін.).

На виході алгоритм дає таку інформацію: вказівка на те, якою частиною мови є аналізоване слово; номер морфеми (рід, число, відмінок – для іменника; рід, число – для дієслова минулого часу; особа, число – для дієслів теперішнього і майбутнього часу.

Питання для самоконтролю

1. Яке місце морфологічного аналізу в процесі автоматичної обробки текстової інформації?
2. Проаналізуйте лінгвістичні засади створення автоматичного морфологічного аналізу (АМА).

3. Чим відрізняються експериментальні та промислові системи АМА?
4. Яка типологія алгоритмів автоматичного морфологічного аналізу?
5. Які принципи роботи аналізаторів АОТ, Mocky, Mystem.

Завдання

1. За допомогою сайту http://starling.rinet.ru/morph.htm виконайте морфологічний аналіз російських слів «кислород», «обезвредил», «научно-исследовательский» та англійських слів «expert», «stimulate» и «favourite». Які можливості пропонує сервіс для російської та англійської мов. Чи зустрічаються помилки? Поясніть причини. Чи є подібні сервіси для української мови? Укладіть їх картотеку.
2. Випишіть із енциклопедії «Українська мова» визначення термінів *граматичне значення, лексичне значення, граматична форма, граматична категорія*, *морфема, морф*, проаналізуйте з позицій комп'ютерної лінгвістики.

ТЕМА 9
МАШИННИЙ ПЕРЕКЛАД ЯК РІЗНОВИД ІНТЕЛЕКТУАЛЬНИХ СИСТЕМ АОТ

План

1. Поняття перекладу як одного з видів мовної діяльності.
2. Види перекладу.
3. Засоби комп'ютеризації процесу перекладу.
4. Поняття машинного перекладу.
5. Поняття перекладацької пам'яті (ТМ).
6. Загальна характеристика засобів автоматизації перекладу.
7. Лінгвістична якість машинного перекладу.

Теоретична частина

Для того, щоб розглядати автоматизований переклад науково-технічної інформації, доцільно спочатку дати визначення та розглянути таке поняття як переклад. Різні вчені по-різному тлумачать це поняття. Це пов'язано у першу чергу із складністю та багатозначність поняття. Тому навіть саме розгорнуте визначення не може охопити усі його істотні властивості.

Термін «переклад «може бути витлумачений у двох напрямках:

1) результат процесу перетворення, тобто сам перекладений текст. Це значення можна виразити через слово друготвір.

2) процес перетворення (перекодування) з мови 1 на мову 2, в результаті якого виникає текст перекладу.

Отже, під час перекладання здійснюється передавання інформації, що міститься у словесному тексті першотвору, у текст друготвору. Існує цілий ряд визначення перекладу як процесу.

За тлумачним перекладознавчим словником поняття «переклад» – це:

1) вид людської діяльності, спрямований на відтворення одиниць мови оригіналу (МО) в мові перекладу (МП) з метою забезпечення комунікації та інформаційного обміну;

2) процес діяльності перекладача по забезпеченню комунікації між носіями різних мов та обміну інформацією між ними;

3) процес міжмовного перетворення або трансформація усного чи письмового тексту однієї мови на іншу;

4) заміна текстового матеріалу на одній мові еквівалентним текстовим матеріалом на іншій;

5) процес переробки інформації в ситуації, коли текст надходить на одній мові, а на виході – на іншій.

Отже, виходячи з цих визначень, можна сказати, що процес перекладу розуміється як трансформація тексту однієї мови на текст іншої. Процес перекладу – це процес перекодування. Це процес обробки, що веде від тексту МО до тексту МП, в основі якого лежить змістове та стилістичне осмислення оригіналу. Ці визначення є досить доцільними, але охоплюють тільки деякі сторони цього складного поняття, та не розкривають всю його багаторівневість.

Відомий лінгвіст Л. Бархударов говорив, що переклад можна вважати певним різновидом трансформації, а саме міжмовної трансформації. Переклад розглядається Л. Бархударовим як деяка трансформація, під час якої зберігається незмінним семантичний інваріант, тобто зміст оригіналу, а форма його вираження, тобто поверхнева структура, може бути змінена».

І. Алексєєва писала, що переклад – це діяльність, яка полягає в варіативному перекодуванні тексту, що був породжений на одній мові, в текст на іншій, що здійснюється перекладачем, який творчо обирає варіант в залежності від варіативних ресурсів мови, видів перекладу, задач перекладу, типів тексту та під впливом власної

індивідуальності.

За словами О. Паршина, переклад в будь-якому випадку представляє собою творчу розумову діяльність, виконання якої вимагає від перекладача цілого комплексу знань, умінь і навичок, здатності робити правильний вибір, враховуючи всю сукупність лінгвістичних та екстралінгвістичних факторів. Переклад – це засіб, який дозволяє забезпечити можливість спілкування (комунікації) між людьми, що говорять на різних мовах.

А. Федоров вважає, що переклад означає вміння висловити вірно і повно засобами однієї мови те, що вже виражено раніше засобами іншої мови.

Л. Черняхівська визначає переклад як перетворення структури мовленнєвого твору, в результаті якого, при збереженні незмінним плану змісту, змінюється план вираження, одна мова замінюється іншим.

М. Зарицький запропонував таке визначення цьому поняттю: це різновид міжмовного спілкування, рецептивно-продуктивної мовленнєвої діяльності. При цьому сприйнятий текст мови-донатора (рецептивний акт) відтворюється мовою друготвору (продуктивний акт). Даний процес можна зобразити у такий спосіб: Т1 (МП) – Т2(МД).

Слід пам’ятати, що процес перекладу залежить як від лінгвістичних так і від екстралінгвістичних чинників. Важливо, також, брати до уваги, що переклад це не лише процес, а ще й результат цього процесу.

Для того, щоб говорити про процесуальні можливості використання засобів автоматизації перекладу в процесі науково-технічного перекладу, треба спершу визначити його місце науково-технічного перекладу та автоматизованого перекладу серед інших видів перекладу. Системний аналіз перекладацької практики дозволяє побудувати єдину типологію перекладу, яка узагальнює різні аспекти підготовки, виконання, представлення, функціонування перекладу.

Типологія будується на таких параметрах:

1. Мова перекладу і мова оригіналу;

2. Перекладач і автор оригінального тексту;

3. Тип перекладацької сегментації та засіб переробки матеріалу, який перекладався;

4. Форма представлення тексту перекладу та форма

представлення оригіналу за формою спілкування;

5. Характер співвідношення тексту перекладу і тексту оригіналу;

6. Жанрово-стилістичні особливості;

7. Тип передачі смислового змісту;

8. Основні функції;

9. Первинність тексту оригіналу;

10. Адекватність;

11. Часові параметри;

12. Ступінь спорідненості.

За параметром перекладач і автор оригінального тексту виділяють такі види перекладу: авторський (тобто переклад виконано самим автором тексту); авторизований (переклад редагує сам автор); машинний (тобто виконана на комп'ютері дія з перетворення тексту однією природною мовою в еквівалентний за змістом текст на іншу мову, а також результат такої дії); змішаний (переклад з використанням значної частки традиційної (чи машинної) переробки тексту).

За параметром форма представлення тексту перекладу та форма представлення оригіналу за формою спілкування виділяють такі види перекладу: 1) письмовий (переклад, виконаний в писемній формі: письмовий переклад писемного тексту (переклад письмового тексту, виконаний у писемній формі); письмовий переклад усного тексту (переклад усного тексту, виконаний у писемній формі)); 2) усний (переклад, виконаний в усній формі: усний переклад усного тексту (переклад усного тексту, виконаний в усній формі); усний переклад письмового тексту (переклад письмового тексту, виконаний в усній формі); послідовний (різновид усного перекладу, здійсненого після прослуховування певної одиниці тексту, в паузах між цими одиницями); однобічний переклад (усний переклад, здійснений тільки в одному напрямку, тобто з однієї мови на будь-яку іншу мову); двосторонній (послідовний усний переклад розмови, здійснений з однієї мови на іншу і навпаки)).

За І. Алексєєвою виокремлюють такі типи перекладу як:

– усний переклад (усний послідовний переклад (абзацний-фразовий переклад): однобічний / двосторонній; синхронний переклад (усний переклад, здійснений практично одночасно з виголошенням тексту-оригіналу); переклад з листа);

– письмовий переклад (машинний (комп'ютерний) переклад; переклад + адаптація (пристосування тексту до рівня компетентності реципієнта; як правило, це спрощення тексту); переклад + стилістична / літературна обробка (відновлення єдності стилю та вирівнювання логіки змісту, оскільки це не в повній мірі вдалося авторові оригіналу); авторизований переклад і співавторство (перекладач вносить власні зміни, міняє сюжет і склад героїв тощо); вибірковий переклад (переклад тільки тієї інформації, яку вимагає замовник)).

За параметром жанрово-стилістичні особливості виділяють такі види перекладу: науково-технічний (переклад науково-технічних текстів і документації); суспільно-політичний (переклад суспільно-політичних текстів); художній (переклад художніх текстів); військовий (переклад текстів з військової тематики); юридичний (переклад текстів юридичного характеру); розмовно-побутового характеру (переклад текстів розмовно-побутового характеру).

Сьогодні вже недостатньо просто перекласти текст, користуючись комп'ютером як друкарською машинкою. Від професійного перекладача очікується, що оформлення готового документа буде відповідати зовнішньому вигляду оригіналу настільки точно, наскільки це можливо, при цьому цей документ повинен задовольняти прийнятим у даній країні стандартам. Від перекладача потрібно також уміння ефективно використовувати раніше виконані переклади на ту ж тему. Очікується, що перекладач буде працювати ефективніше, швидше та з помітною економію коштів. Ці жорсткі, найчастіше суперечливі умови можна дотримати лише в тому випадку, якщо перекладач не тільки досконало володіє рідною та іноземною мовою і глибоко вивчив обрану ним предметну область, а й впевнено орієнтується в сучасних комп'ютерних технологіях.

Незважаючи на те, що програми, які оснащені пам'яттю перекладу, називаються системами автоматизованого перекладу (CAT, computeraided/assisted translation), їх не слід плутати з програмами машинного перекладу (machine translation) – пам'ять перекладу нічого не перекладає сама по собі, в той час як машинний переклад заснований на генерації переказів за результатами граматичного розбору вихідного тексту. Машинний переклад – це спроби лінгвістів, програмістів і фахівців з штучного

інтелекту створити таку програму, яка могла б повністю замінити перекладача, але домогтися цього – вкрай складне завдання. Зрозуміло, що при нинішньому, не досить високому, рівні машинного перекладу без участі людини не обійтися. Щоб комп'ютер міг перекласти текст, йому потрібна допомога передредактора, який тим чи іншим чином попередньо обробляє текст для перекладу, інтерредактора, який бере участь в процесі перекладу, і постредактора, який виправляє помилки і недоліки в перекладеному машиною тексті. Машинній переклад – це автоматичний процес перекладу з однієї природної мови на іншу спеціальною комп'ютерною програмою. Ці комп'ютерні програми засновані корпусах текстів. Такі сучасні комп'ютерні програми перекладу є досить ефективними для використання, але вони досі не можуть вирішити одну з найгостріших проблем процесу перекладу: вибір контекстуально потрібного варіанту перекладу, котрий обумовлений багатьма причинами у кожному тексті. Зараз результат цього виду перекладу може бути використаний як чернетка майбутнього тексту, який вимагає редакції перекладача.

Виникає питання, що являє собою технологія перекладацької пам'яті (ТМ), та чому вона є більш ефективною під час перекладу науково-технічної літератури. Отже, у основі всіх систем автоматизації перекладу лежить технологія ТМ. ТМ – це база даних, де зберігаються виконані переклади. Технологія ТМ працює за принципом накопичення: в процесі перекладу в ТМ зберігається вихідний сегмент (пропозиція) і його переклад. При обробці нового тексту, що надійшов для перекладу, система порівнює кожне його речення зі збереженими в базі сегментами. Якщо ідентичний або подібний вихідному сегмент знайдений, то переклад цього сегмента відображається разом з перекладом і вказівкою збігу у відсотках. Слова та фрази, які відрізняються від збереженого тексту, виділяються підсвічуванням. Таким чином, перекладачеві залишається перекласти тільки нові сегменти і відредагувати ті, що частково збігаються. Кожна зміна або новий переклад зберігаються в ТМ. Отже, в результаті немає необхідності двічі перекладати одне і те ж саме речення. З іншого боку, при роботі з великими проектами перекладач зустрічається з проблемою узгодженого застосування термінологічного глосарію в ході тривалого проекту або швидкого повторного використання раніше перекладеного тексту. За своєю природою подібні рутинні завдання порівняно

легко (на відміну від машинного перекладу) формалізуються і програмуються. Кожен запис бази даних ТМ являє собою одиницю (речення або абзац) паралельних текстів (як правило, двома мовами). За словами В. Широкова, мовознавця, що працював в галузі лексикографії, це стало можливим завдяки розвитку керованих лексикографічних баз на основі теорій та практики корпусної лінгвістики. Така база даних зберігає попередні переклади з метою їх можливого повторного використання і вирішення завдань швидкого пошуку по вмісту. Як правило, запис пам'яті перекладу складається з двох сегментів: на вхідній і кінцевій мові. Якщо ідентичний (або схожий) сегмент МО зустрічається в тексті, сегмент МП буде знайдено в пам'яті перекладу і запропоновано перекладачеві як основу для нового перекладу. Автоматично знайдений текст може бути задіяний так, як його запропонувала програма, відредагований або повністю відкинутий. Більшість програм використовують алгоритм нечіткої відповідності. Побудова моделей наближених роздумів людини і використання їх у комп'ютерних системах представляє сьогодні одну з найважливіших проблем науки. Основи нечіткої логіки були закладені наприкінці 60-х років у працях відомого американського математика Латфі Заде. Першим серйозним кроком у цьому напрямку з'явилася теорія нечітких множин, розроблена Заде. Його робота «Fuzzy Sets», що з'явилася в 1965 році в журналі «Information and Control», заклала основи моделювання інтелектуальної діяльності людини і з'явилася початковим поштовхом до розвитку нової математичної теорії. Отже, за Заде теорія нечітких множин є апаратом формалізації одного з видів невизначеності, що виникає при моделюванні реальних об'єктів. Нечіткість завжди виникає, коли ми використовуємо слова природної мови для опису об'єкта. Що істотно поліпшує їх функціональні можливості, оскільки в цьому випадку можна знаходити речення, що лише віддалено нагадують фрази, які шукає програма, але, тим не менш, вони є придатними для подальшого редагування. Переваги від використання такого програмного забезпечення спочатку можуть бути неочевидні – проте по міру наповнення бази даних результати автоматичної підстановки основ для перекладу будуть ставати все більш точними і регулярними.

Основою систем автоматизації перекладу є технологія ТМ, що є дуже ефективною для перекладача, який переважно працює з

науково-технічними текстами однієї тематики. Системи автоматизації перекладу сприяють біль ефективному, швидкому та якісному перекладу, у зв’язку зі своїми чотирма особливостями функціонування:

1. Ці системи поділяють текст для перекладу на певні сегменти (розміри та особливості сегментів задаються у конфігурації програм), зберігають їх на МО та МП, а потім видають ці сегменти у зручній для перекладача формі, для того, щоб зробити процес перекладу більш зручним та швидшим.

2. Так як ці сегменти на МО та МП зберігаються програмою, та являють собою одиниці перекладу, перекладач у будь-який момент може повернутися до них та перевірити їх переклад. Для цього системи автоматизації перекладу зазвичай оснащені певними інструментами, що допомагають швидко орієнтуватися в тексті, знаходити потрібні вам сегменти для перекладу чи перевірки якості перекладу.

3. Сегменти, що виділяються програмою, зберігаються в її пам’яті в спеціальних перекладацьких базах даних, отже вони завжди можуть бути використані під час перекладу.

4. І остання особливість роботи цих програм полягає в автоматичному пошуку сегментів у базі даних, їх відображення та вставці результатів пошуку в текст перекладу. Архітектура автоматизованої системи та її функціональні можливості можуть відрізнятися. Інструменти пошуку можуть працювати як з цілими сегментами, так і з окремими словами чи фразами, дозволяючи перекладачеві виконувати термінологічний пошук. У систему також включають окрему програму для роботи з глосарієм, що містить затверджені для застосування в терміни. Деякі системи працюють з програмами машинного перекладу. Основний робочий інтерфейс або вбудовується безпосередньо в наявний текстовий процесор, такий як Word, або є окремий редактор. До складу системи обов’язково включають фільтри для імпорту-експорту файлів різних форматів. Крім того, багато систем мають засіб для додавання в пам'ять перекладу сегментів з наявних у перекладача старих перекладених файлів.

Сучасний розвиток усіх сфер людської життєдіяльності ставить нові завдання в комунікаційному просторі людства. Коли рух інформаційних потоків не знає меж у часі і просторі, роль перекладу невпинно зростає. До традиційних видів останнього –

письмовий, усний, синхронний, послідовний, з листа та ін. – додається машинний переклад, час практичного використання якого розпочався з 1980-х років. У наш час є достатньо широкий вибір програм, які полегшують працю перекладача, котрі умовно можна підрозділити на дві основні групи: електроні словники (electronic dictionary) та системи машинного перекладу (machine translation system). Системи машинного перекладу забезпечують послідовний переклад текстів, що враховує морфологічні, синтаксичні та семантичні зв'язки членів речення. Програми перекладу (системи машинного перекладу) з'явилися у відповідь на потреби користувачів в оперативному перекладі різної комерційної, технічної або Інтернет – інформації, яка подана в електронному вигляді. Аналізуючи програми машинного перекладу, потрібно відразу зазначити, що вимоги до них не повинні бути такими ж, як і до перекладу, який виконує людина. Переклад, зроблений комп'ютером, поки що далеко не ідеальний, але текст, отриманий в результаті роботи електронного перекладача, дозволяє у більшості випадків зрозуміти суть документа, який перекладається. Далі цей документ можна коригувати, маючи базові знання іноземної мови та добре орієнтуючись в предметній галузі, до якої належить інформація, що перекладається.

Вперше можливість машинного перекладу на практиці передбачив Ч. Беббідж, що у першій половині 19 століття працював над проектом цифрової аналітичної машини – механічного прототипу електронних цифрових обчислювальних машин. Процес перекладу він уявляв так: «В мене перед очима текст, написаний російською, але я збираюсь уявити, що насправді він написаний англійською, але за допомогою доволі дивних знаків. Все, що мені потрібно – це зламати код для того, щоб вилучити інформацію, що міститься у тексті». Перші патенти на створення перекладацьких машин було видано у середині 30-х років минулого століття. Ідея науковця з Західної Європи полягала у створенні автоматичного двомовного словника на основі перфострічки, але проект росіянина П. Троянського був детальнішим. Винайдений ним пристрій включав двомовний словник, здатний оперувати граматичними особливостями за принципом мови есперанто. Система поділялася на три стадії. Спочатку носій мови мав розподілити слова за їх логічними формами та синтаксичними функціями. Потім машина мала виконати переклад на потрібну мову, а носій – відредагувати

висхідний матеріал.

Першу пропозицію машинного перекладу за допомогою комп'ютера було висунуто Уорреном Вівером, дослідником з Фонду Рокфеллера у його меморандумі. Пропозиції базувалися на інформаційній теорії, успіхах у зламуванні кодів протягом другої світової війни та обговореннях універсальних та основних принципів мов. За кілька років після опублікування меморандуму розпочалися серйозні дослідження у багатьох університетах Сполучених Штатів. 7 січня 1954 у Нью-Йорку в головному офісі IBM було вперше проведено публічну демонстрацію системи машинного перекладу (МП). Про демонстрацію повідомили в газетах, тож подія отримала широкий розголос. Попри те, що сама система мала лише 250 слів та 49 перекладених на англійську російськомовних речень (головним чином у області хімії) і була доволі примітивною, вона продемонструвала перспективи машинного перекладу, стимулювавши фінансування цього дослідження не тільки у США, а й у всьому світі. Експеримент було визнано успішним, що сповістило про початок ери вагомих капіталовкладень у дослідження машинного перекладу. Автори стверджували, що за кілька років машинний переклад буде повністю втілено в життя.

У ранніх системах використовувались великі двомовні словники та закодовані вручну правила для визначення порядку слів у висхідному продукті. У результаті цей метод було визнано занадто обмеженим, а завдяки тогочасному розвитку лінгвістики для покращення якості перекладу було запропоновано дослідження генеративної лінгвістики та трансформаційної граматики. Але в цей час операційні системи вже застосовувались. Військово-повітряні сили США використовували систему, розроблену IBM і Вашингтонським університетом, в той час як на ВПС Італії працювала розробка Джорджтаунського університету. Попри те, що якість продукції була низькою, це задовольняло клієнтів, в основному з точки зору швидкості.

Наприкінці 1950-х Г. Бар, дослідник, що на замовлення США вивчав можливість створення повністю автоматичного якісного перекладу, наголосив на проблемі семантичної двозначності під час машинного перекладу. Розгляньмо наступний приклад: Little John was looking for his toy box. Finally he found it. The box was in the pen. Слово «pen» має два значення: прилад, що використовується на

письмі і певний контейнер. Для людини значення є очевидним, але машина без «універсальної енциклопедії» ніколи не зможе вирішити цю проблему. Сьогодні проблема семантичної двозначності може бути вирішена шляхом написання висхідних текстів контрольованою мовою, тобто застосовуючи словник, у якому для кожного слова є тільки одне значення. Дослідження 1960-х років у СРСР та Сполучених Штатах сконцентрувались головним чином на російсько-англійській мовній парі. В основному об'єктами перекладу виступали науково-технічні документи, як от статті з наукових журналів. Недбалий переклад був достатнім для розуміння сенсу статей. Якщо стаття стосувалася інтересів безпеки, її надсилали живому перекладачу для повного перекладу, решту ж перекладали автоматично. МП зазнав нищівного удару у 1966 році разом із публікацією звіту ALPAC (дорадчого комітету з автоматичної мовної обробки), що складався з семи вчених, скликаних американським урядом у 1964. Американський уряд був занепокоєний повільним просуванням експерименту попри значні видатки, тож було ухвалено рішення, що машинний переклад був дорожчим, менш точним та повільнішим за людський, і незважаючи на витрати, машинний переклад навряд чи досягне якості людського найближчим часом.Однак, у звіті рекомендували продовжувати дослідження у галузі комп'ютерної лінгвістики та створити автоматичні словники, що допомагали б перекладачам. Публікація звіту вплинула на дослідження машинного перекладу у Сполучених Штатах, меншою мірою – у Радянському Союзі та Великій Британії. Звіт, принаймні, майже повністю припинив будь-які дослідження у Сполучених Штатах майже на десятиліття. Однак, у Канаді, Франції та Німеччині дослідження продовжувались; у 1970 систему Systran використовували ВПС США, і, як наслідок, Комісією європейської економічної спільноти. Систему METEO, розроблену в Університеті Монреаля, було застосовано у 1977 року в Канаді для перекладу прогнозів погоди з англійської мови на французьку. Система перекладала близько 80000 слів в день або 30 млн слів на рік, поки не була замінена системою конкурента 30-го вересня 2001. В той час як дослідження 1960-х концентрувалися на проблемі обмеження мовних пар та системі вводу інформації, протягом 1970-х років постала потреба в недорогих системах, що змогли б перекладати діапазон технічних та комерційних документів. Ця вимога заохочувалася посиленням

глобалізації та потреби перекладів в Канаді, Європі та Японії. До 1980-х років збільшилось різноманіття та число встановлених систем для машинного перекладу. Збільшилась кількість систем, що працювали на основі електронно-обчислювальних машин, як от Systran і Logos. Поширення мікрокомп'ютерів знаменувало створення дешевого ринку систем машинного перекладу, тому багато європейських, американських та японських компаній не втратили шансів цим скористатися. Схожі системи потрапили на ринки Китаю, Східної Європи, Кореї та Радянського Союзу. Протягом 1980-х жвава діяльність у галузі машинного перекладу розгорнулася в Японії. Завдяку комп'ютеру п'ятого покоління Японія мала намір перестрибнути конкурентів у галузі елементів електронних пристроїв та програмного забезпечення. Багато великих японських компаній було залучено до роботи над створенням англо-японських та японо-англійських штучних перекладачів. У дослідженнях 1980-х років переклад вбачався як певний різновид проміжного лінгвістичного відтворення, що залучає морфологічний аналіз разом з синтаксичним та семантичним. Наприкінці 1980-х відбувся вагомий стрибок у галузі створення нових методів для машинного перекладу. Система, розроблена IBM, базувалася на статистичних методах, інші групи застосовували техніку, що базується на великій кількості перекладів у якості зразків.

Протягом 1990-х років, після успіхів у розпізнанні мови та мовному синтезі, дослідження перейшло у стадію мовного перекладу. Відбувся значний ріст використання машинного перекладу в результаті появи дешевих та потужних комп'ютерів. На початку 1990-х машинний переклад став можливим не тільки на великих базових комп'ютерах, а й на персональних комп'ютерах та автоматизованих робочих місцях.

Таким чином, з перших же кроків у машинному перекладі вималювалися три основні інваріантні функціонально-процедурні блоки: 1) аналіз вхідного тексту за допомогою спеціального машинного словника та граматики; 2) перетворення результатів аналізу на інформацію, необхідну для побудови перекладеного еквівалента; 3) синтез вихідного тексту, що використовує два типи інформації: отриману на етапі трансферу та граматику синтезу вихідного тексту.

Машинний переклад еволюціонував від максимально

спрощених версій («лексиконних») до версій, «заглиблених» у зміст. Його еволюцію можна представити як «естафету» п'яти послідовних поколінь та, відповідно, типів систем:

1. Системи послівного перекладу (word-for-word translation), доповнені деякими допоміжними граматичними конструкціями – системи, що здатні опрацьовувати лише окремі мовні ситуації в конкретних випадках, але не здатні оперувати ними в цілому, тобто на рівні типів, класів, груп тощо.

2. Структурно-граматичні системи машинного перекладу, що базуються на морфологічних кореляціях між вхідною та вихідною мовами – «морфологічні системи» – виявляються ефективними для організації перекладу в межах споріднених мов (наприклад, українсько-російський машинний переклад та навпаки).

3. Структурно-граматичні системи машинного перекладу, що спираються на синтаксичні кореляції між вхідною та вихідною мовами – «синтаксичні системи». Центральною процедурою тут є синтаксичний аналіз вхідної фрази, яка далі трансформується в структурно-синтаксичний каркас вихідної фрази.

4. Структурно-семантичні системи машинного перекладу, що оперують глибинними структурами вхідного та вихідного контекстів. У таких системах передбачається багаторівневе опрацювання мовного матеріалу.

5. Автоматизовані робочі місця (станції) перекладача – інтерактивні словниково-орієнтовані системи машинного перекладу з великими за обсягом і деталізованими термінологічними словниками. Концепція полягає в тому, що системи машинного перекладу принципово неспроможні забезпечити його високу якість, тому вони не можуть цілком замінити користувача-перекладача, але мають допомагати йому.

Отже, наприкінці XX сторіччя машинний переклад існує вже не лише як теоретична дисципліна з експериментальним супроводом, але і як технологічна реальність. Проте, навіть при такому рівні розвитку сучасних систем автоматичного перекладу, не завжди вдається правильно перекласти деякі словосполучення у сферах професійного спрямування.

При створенні СМП, які базуються на використанні лінгвістичних правил, потрібне знання розпізнавання ознак тексту, що відносяться до сфери прагматики: жанр та стиль (наприклад, це публіцистична стаття, вірш чи документ встановленого зразка);

область знання, до якого текст відноситься (розпізнавання термінології); зв’язаність частин тексту, що не завжди описується за допомогою синтаксичних чи лексико-семантичних критеріїв; і т.д.

Можна вважати, що для адекватного перекладу автоматична система повинна: знати внутрішні структури мов, між якими здійснюється переклад; мати ясне уявлення про культуру, історію, мораль, переважні типи мислення народів, що є носіями мови; володіти по можливості більшим словниковим запасом, більш-менш структурованим по областях застосування слів (спеціальна термінологія, діалекти, ідіоматика, сленг); мати явний чи інтуїтивний тезаурус слів обох мов, тобто по даному слову вміти запропонувати семантичні функції від нього, такі як синонім, антонім, конверсив, класичний атрибут, а також уміти запропонувати похідні частини мови від даного слова, якщо такі існують (добро – добрий – добріше – подобрів і т.п.).

Використання систем машинного перекладу (СМП) з українською та російською мовами є наразі дуже поширеним в Україні. Проте, якщо при перекладі тексту з російської на українську мову користувачеві пропонуються скальковані російські форми з автоматичного словника. Отже, питання покращення таких словників, на наш погляд, є досить важливим. Інакше переклад, який ряснітиме покручами, не матиме жодної цінності. Вичитуючи перекладений з російської мови текст, пересічний користувач без відповідної філологічної підготовки виправляє, як правило, звичайні помилки, пов’язані з узгодженням слів, відмінковим та прийменниковим керуванням. При цьому лексичні помилки часто залишаються непоміченими, оскільки, на думку користувача, якщо слова є в автоматичному словнику, то вони мають бути правильними. Проблема полягає в тому, що навіть академічні словники містять подекуди форми, скальковані з російської мови.

Для оцінювання цього напряму перекладу використовують такі системи машинного перекладу: «Ленгвідж Майстер 98», «Прагма 5.0» та «Плай 5.0» (надалі «ЛМ», «Прагма», «Плай»). При перекладі текстів системами машинного перекладу часто трапляються слова, які відсутні в автоматичному словнику. Як правило, це терміни певної галузі або власні назви. Неперекладене слово суттєво впливає на загальну якість перекладу, оскільки воно випадає із загального аналізу й може призвести до неправильного

синтезу вихідного речення. Кількість неперекладених слів також може свідчить про якість системи.

Навіть у таких близьких мовах, як українська та російська, моделі словотворення збігаються не завжди. За цілком слушною думкою С. Караванського, «... кожна мова творить у процесі розвитку свою власну модель парування прикметників з іменниками, і ця модель не може око-в-око повторитися в іншій мові». Так, наприклад, для перекладу російського слова «жесткий» українці вживають чотири різні прикметники: твердий (вагон), круті (заходи), тяжкі (умови), суворий (контроль). Для перекладу слова «жесткий» СМП «Прагма» використовує лише одне слово – «жорсткий». У словнику системи «Плай» є ще один варіант перекладу – «твердий», що збільшує можливість правильного перекладу, але не відповідає всім варіантам вжитку слова «жесткий». Кожна мова в процесі свого розвитку намагається уникати довжелезних і складних конструкцій. Будь-який мовець з розвиненим чуттям мови не вживатиме таких форм. Короткі та природні форми є окрасою будь-якої мови, вони створюють її неповторну палітру й полегшують спілкування. Так, наприклад, кожна з трьох систем машинного перекладу перекладає російське словосполучення «должны были» як «повинні були». А якщо речення має складний присудок, то взагалі отримуємо занадто складну форму: «... резолюції, які повинні були бути включені в загальну постановку» («ЛМ»). Звичайно, що форма «мали бути включені» буде більш віповідною й звучатиме по-українському. З трьох систем, які ми оцінюємо в цій статті, найкраще перекладає керування дієслів «Прагма», хоча результати перекладу цього граматичного явища в кожній системі досить невтішні. Крім того, до грубих помилок належать випадки хибного перекладу лише одного слова, переклад якого не потребує будь- якого аналізу та жодних перекладацьких трансформацій. Наведемо приклади таких помилок: «болільник» замість «вболівальник» («Плай»), «бастували» замість «страйкували» («Прагма»). Схожі помилки мають бути виправлені в автоматичних словниках у першу чергу. Їх виправлення ніяк не пов’язане зі зміною алгоритму й не може призвести до погіршення якості перекладу. Серйозною проблемою для багатьох мовців є переклад на українську мову російських дієприкметників. Як відомо, з погляду стилістики української мови краще уникати частого вживання дієприкметників. Коли йдеться

про активні та пасивні дієприкметники теперішнього часу, то вони практично відсутні в сучасній українській мові, що є її лексико-граматичною особливістю, а не вадою. Тестованим нами системам це явище, здається, невідоме. Майже всі російські дієприкметники з добірки тестових речень були перекладені тільки дієприкметниками. Замість них можна вживати прикметники: «панівні кола», а не «правлячі» («Плай»); «життєствердні мотиви», а не «життєстверджуючі» («ЛМ»); «всеохопний підхід, а не «усеохоплюючий» («Плай»). Досить продуктивною є заміна активних дієприкметників іменниками: не «командуючий» («Прагма»), а «командувач флотом»; не «завідуючий кафедрою» («ЛМ»), а «завідувач кафедри».

Системи машинного перекладу третього покоління володіють достатнім інструментарієм для адекватного перекладу сталих словосполучень. Проте оцінювання та аналіз тестованих речень для російсько-українського напряму свідчить про те, що з цим видом у перекладацьких систем поки, на жаль, не все гаразд. Найбільшу групу сталих словосполучень становлять дієслівно-іменникові словосполучення, частина яких перекладена цілком пристойно: «Суперник чинив гідний опір» («Плай»); «Питання, звичайно, не однозначне, але в цьому він, безумовно, має рацію» («Прагма»); «З метою безпеки, охорона була вимушена вжити найсуворіших заходів» («ЛМ»). Проте більшість словосполучень перекладено невдало: «відволікати увагу» («Прагма») замість «відвертати увагу»; «взяти напрокат автомобіль» («Плай») замість «випозичати авто». Таким чином, робота над покращенням якості систем автоматичного оброблення текстової інформації, зокрема систем машинного перекладу, потребує багато зусиль та часу, але це не безплідна робота.

Отже, основною прерогативою роботи розробників СМП у межах цієї мовної пари є наразі, на наш погляд, поліпшення автоматичних словників. Оскільки досягнення ідеальної якості машинного перекладу – завдання надто складне, то переклад може бути недосконалим, але принаймні українським, а не калькованим, щоб запобігти засміченню та дискредитації мови.

Питання для самоконтролю

1. Назвіть основні віхи історії машинного перекладу.
2. Які є принципи машинного перекладу?

3. В чому полягає лінгвістичне забезпечення систем машинного перекладу?

4. Яка структура системи машинного перекладу?

5. Що таке системи машинного перекладу та яка їх класифікація?

6. Що таке системи прямого перекладу?

7. Що таке системи перекладу за принципом трансферу?

8. Що таке системи перекладу з мовою-посередником?

9. Назвіть алгоритми машинного перекладу.

10. В чому специфіка моделі машинного перекладу на основі перекладних відповідників АМПАР?

11. Які особливості система машинного перекладу СИСТРАН?

12. Що таке джорджтаунська система машинного перекладу?

13. Які особливості системи РУТА?

14. Які типові помилки машинного перекладу?

15. Назвіть основні проблеми комп'ютерної лексикографії?

16. Які вимоги висуваються до систем машинного перекладу?

Завдання

1. Використовуючи список літератури до теми та сучасні статті науковців, розміщені у мережі Інтернет, напишіть реферат та підготуйтеся до його захисту із застосуванням презентації (реферат оформити відповідно до Положення з написання студентських наукових робіт у вищій школі).

2. Відскануйте та розпізнайте уривок статті (близько 150 слів) за допомогою програми ABBYY FineReader. Перекладіть цей текст за допомогою систем МП Pragma та «Google Переводчик». Відредагуйте перекладені тексти, письмово проаналізуйте та опишіть помилки систем Pragma та «Google Переводчик».

ТЕМА 10
КОМП'ЮТЕРНІ ТЕХНОЛОГІЇ ПІДГОТОВКИ ТЕКСТОВИХ ДОКУМЕНТІВ

План

1. Класифікація програм підготовки текстів.

2. Етапи підготовки текстових документів.

3. Використання систем підготовки текстових документів для виконання складних задач.

4. Тенденції розвитку систем підготовки текстових документів.

Теоретична частина

Обробка текстів як напрям розвитку техніки виник у першій декаді XX ст. із появою механічної друкарської машинки. Потім більше півстоліття машинка лишалась єдиним загальнодоступним засобом одержання друкованого тексту на папері. Найбільш трудомістким процесом було внесення змін у текст. Перший революційний крок у галузі обробки текстів був зроблений фірмою IBM в 1964 р., коли вона випустила систему під назвою MT/ST (Magnetic Tape / Selectric Typewriter) – друкарську машинку із записувальним пристроєм, який дозволяв записувати текст із друкарської машинки на касету з магнітною стрічкою, після чого можна було знайти в тексті потрібне місце, вставити корекцію, знищити частину тексту чи повторити частину тексту без повторного вводу з клавіатури. Пізніше магнітна стрічка була замінена магнітними картами, кожна з котрих містила сторінку тексту і була зручніша, ніж магнітна стрічка, для зберігання і пошуку тексту.

На початку 70-х рр. фірми Lexitron і 3M розробили текстові процесори з відеодисплеями, які дозволяли бачити текст, що вводиться з клавіатури на екрані і вносити зміни, які одразу ж відображались на екрані. У 1973 р. текстові процесори вже мали пристрої запису тексту на гнучких дисках, що дозволяли мати прямий (не послідовний, як на магнітній стрічці) доступ до будь-якої частини тексту. Швидкість роботи істотно зросла. Перші електронні текстові процесори були громіздкими та дорогими, проте з появою мікропроцесора і персональних комп’ютерів на їх основі текстові процесори стали широко доступними. У 80-і рр. було розроблено велику кількість текстових процесорів для різноманітних персональних комп’ютерів, що відрізнялись як функціональними можливостями, так і інтерфейсом користувача. Сьогодні поширені текстові процесори, які можна вважати настільними видавничими системами, що дозволяють виконувати не тільки ввід і редагування тексту, а і верстку в інтерактивному режимі складного тексту з ілюстраціями.

Системи підготовки текстових документів сьогодні значно відрізняються одна від іншої характеристиками, можливостями

вводу і редагування текстів, його форматуванню і виводу на друк, а також за ступенем складності опанування користувачем. При цьому всі текстові процесори використовують різні метафори для побудови документів і різні способи представлення документа для огляду під час роботи над ним. У Word Perfect, текстовому процесорі з найдовшою історією, до сьогодні помітні риси, що відображають його найважливішу мету – зробити користування класичними комп'ютерами 1980-х рр. таким же простим, як друкарською машинкою. Складні документи в середовищі Word Perfect представлені у вигляді простих файлів з розміткою.

На відміну від цього в Microsoft Word for Windows завжди передбачалось, що документи мають структуру і що структурними блоками документа є абзаци та розділи, кожний із яких може бути представленим у власному форматі. Структура у вигляді абзаців і розділів пронизує кожний файл Word. Word Pro увів першу нову метафору організації документа після того, як у 1983 р. у Word з'явилися таблиці стилів. У Word Pro документ поділяється на розділи, які, у свою чергу, можна поділити на підрозділи, змінити порядок і переставити шляхом простого переносу закладок, що факультативно відображаються зверху вікна редагування. Цей тип структури значно гнучкіший, ніж порівняно жорсткі структури у Word. Маніпуляції з елементами структури документа в середовищі Word Pro досить прості для користувача.

Хоча документи Word Perfect мають найпростішу базову структуру, обробляти складні багатосторінкові документи в Word Perfect значно зручніше, ніж у програмах, яким притаманна більша структурованість.

Word Perfect позволяє користувачу встановлювати перехресні зв'язки між нумерацією абзаців і розділів та складовою нумерацією томів, розділів, сторінок. Найслабші можливості в цьому плані має Word Pro, а Word займає проміжне положення.

Вибір конкретного програмного продукту для обробки тексту є дуже відповідальним моментом. Процес вибору пов'язаний із багатьма факторами, але передусім слід керуватись принципом розумної достатності. Найважливішою для практичного користувача характеристикою програми цього класу могла б виступати галузь професійної діяльності, для якої програмний продукт зручний у користуванні. Інструментальні засоби підготовки текстових документів використовуються для набору

текстів програм, документів різного ступеня складності, наукових статей, книг тощо. Обмежувальним фактором може виступати кваліфікація користувача.

Найактуальнішим при описі процесу редагування є поняття інтерфейсу користувача, до якого, передусім, відносять мову спілкування з текстовим процесором, а також пристрій вводу-виводу (для ПК це клавіатура, маніпулятор типа «миша», екран дисплея і принтер). Найпривабливішими для розробника документа є ті програмні середовища, у яких можлива настойка інтерфейсу під свої смаки і потреби. Мова спілкування користувача з комп'ютерною системою підготовки тексту організований на основі діалогових процедур. У стадії розробки знаходяться пристрої мовленнєвого вводу інформації, що дозволяють перетворювати вимовлені слова в текст. Тенденція до включення мовленнєвих компонентів у інтерфейси користувачів у майбутньому посилиться. Розпізнавання мовлення навряд чи колись повністю замінить інші способи вводу, але в перспективі інтерфейси будуть наділеними комбінованими засобами вводу. Ця концепція має назву багаторежимного (multimodal) вводу.

Щодо зручності для користувача однією з найважливіших властивостей текстових процесорів є повна відповідність твердої копії зразку документа на екрані. Така характеристика англійською називається WYSIWYG (What you see is what you get – «що ви бачите, то й отримаєте»).

Серед систем підготовки текстових документів (СПТД) природними мовами можна виділити три класи: форматери, текстові процесори та настільні видавництва. Форматер – система підготовки текстів, яка не використовує для внутрішнього представлення тексту ніяких кодів, крім стандартних: кінець рядка, переведення каретки, кінець сторінки. Текстовий редактор – це система підготовки текстів, яка у внутрішньому представленні постачає власне текст спеціальними кодами – розміткою.

Текстові процесори мають спеціальні функції, які призначені для полегшення вводу тексту і представлення його в роздрукованому вигляді, зокрема: увід тексту під контролем функцій форматування, що забезпечують негайну зміну вигляду сторінки тексту на екрані і розташування слів на ній, даючи уявлення про розташування тексту на папері після друку; можливість попереднього опису структури майбутнього документа

(величина абзацних відступів, тип і розмір шрифту для різних елементів тексту, розташування заголовків, міжрядкова відстань, кількість копій документа, розташування і спосіб нумерування посилань (у кінці тексту чи на тій самій сторінці) тощо); автоматична перевірка орфографії і одержання підказки при виборі синонімів; увід і редагування таблиць і формул з відображенням їх на екрані в тому вигляді, у якому вони будуть роздруковані; об'єднання документів у процесі підготовки тексту до друку; автоматичне складання заголовку і алфавітного довідника.

Із метою сумісності текстових документів при перенесенні їх із середовища одного текстового процесора в інший існує особливий вид програмного забезпечення – конвертери. Сьогодні автономні програми-конвертери практично не існують і стали складовою частиною системи підготовки текстів.

Настільні видавництва готують тексти за правилами поліграфії і з типографською якістю. Вони призначені не стільки для створення великих документів, скільки для реалізації різного роду поліграфічних ефектів, тобто програма настільного видавництва дозволяє легко маніпулювати текстом, змінювати формати сторінок, розмір відступів, дає можливість комбінувати різні шрифти. За функціональними можливостями пакети НВС аналогічні кращим текстовим процесорам, і межа, що їх роз'єднує, стає непомітнішою. Так, обидва типи програм дозволяють розміщувати на одній сторінці текст і ілюстрації, обидва типи дозволяють форматувати текст у декілька колонок, редагувати текст і маніпулювати текстовими блоками. Текстові процесори є простішими за пакети НВС, а НВС мають ширші можливості управління підготовкою тексту. По-друге, підготовлені в НВС матеріали виглядають виданнями вищого рівня якості, а не просто як гарні роздруківки. Всі пакети мають характеристики, відсутні в більшості текстових процесорів, наприклад, стиснення і розтягнення рядків, обертання тексту і зміна відстаней між рядками і абзацами з дуже маленьким кроком приросту тощо.

Зовнішній файл, підготовлений текстовим процесором, можна роздрукувати тільки цим же текстовим процесором. Як правило, друк може виконуватись на принтері будь-якого типу. Тексти, підготовлені настільними видавництвами, роздруковуються тільки на лазерних принтерах. Серед систем підготовки текстових документів у цьому класі виділяють настільні видавництва

професійного рівня і видавничі системи початкового рівня. Системи першої групи призначені для роботи над видавництвами документів зі складною структурою чи типу ілюстрованого журналу. До систем професійного рівня можна віднести QuarkXPress for Windows, FrameMarker for Windows, PageMarker for Windows. Опанування дорогий і складних в експлуатації «настільних типографій» зазвичай вимагає значних часових затрат, тому їх недоцільно використовувати спеціалістам, які за родом занять лише деколи мають гарно і швидко підготувати документацію, лист чи оголошення.

Системи другої групи не призначені для друку промислової поліграфічної продукції. Користувачі цього класу НВС для вирішення своїх задач, як правило, використовують інші програми, НВС використовують епізодично, наприклад, для створення інформаційного бюлетеня чи форматування вітальної листівки для тиражування в невеликій фірмі. Всі пакети даної категорії орієнтовані на новачка і користувача, який займається видавничою діяльністю зрідка. Найрозповсюдженіші в цій групі Microsoft Publisher, Pageplus for Windows. Найочевиднішим параметром оцінки рівня НВС є її ціна.

На ринку постійно з’являються нові версії редакторів із більш досконалими і складними процедурами обробки, що призводить до розмивання меж між класами.

Підготовка текстів із використанням систем підготовки текстових документів (СПТД) полягає в послідовному виконанні таких етапів: набір тексту; редагування введеної інформації; форматування (оформлення) окремих структурних елементів майбутнього документа; друк документа; збереження тексту документа та ведення архіву текстів.

При роботі з текстом зазвичай відбувається чергування операцій різних етапів, тому окремі операції не можна чітко віднести до певного етапу підготовки документа.

Набір тексту здійснюється на клавіатурі. Розміри службової та робочої частин екрана дисплея залежать від конкретної програми підготовки тексту. Звичайно в службовій частині присутній рядок статусу, який містить корисну для користувача інформацію про режим роботи програми підготовки тексту і використовується для короткої підказки про дію програми при виборі того чи іншого пункту меню. Будь-яка система підготовки текстів підтримує два

режими вводу – вставки чи заміни. Клавіша <Backspase> знищує помилково набраний символ ліворуч від курсора, клавіша <Delete> знищує з тексту символ, що знаходиться в позиції курсору, а всі символи праворуч зміщуються ліворуч у режимі вставлення. Прокрутка чи скролінг рядків – це процес автоматичного зсуву тексту, коли верхній рядок зникає з поля зору, а знизу з'являється порожній рядок для подальшого введення символів.

До часто порушуваних правил комп'ютерного набору відносять наступні: розділові знаки відбиваються від попереднього тексту; допускається лише відбивати пробілом лише знак питання; не відбиваються від цифр знаки процент, градус, хвилина, секунда (45%, 20°); одним пробілом цифри відбиваються від №, але при введенні слід використовувати «нерозривний» пробіл, який не буде ані розтягнутим при форматуванні, ані розірваний при переносі; дефіс пробілами не відбивається; в комп'ютерному наборі часто неправильно використовують дефіс замість тире, тому що знака тире в стандартній розкладці клавіатури нема; тире, навпаки, відбивається пробілом; тире не відбивають пробілом у сполученнях, що мають значення «від» і «до», наприклад: 1941–1945 рр.

Пам'ятка: не можна закінчувати рядок прийменником чи сполучником, із якого починається нове речення.

Часто порушувані правила використання скорочень: у кінці загальноприйнятих скорочень кг, т, ц, км крапка не ставиться; скорочення типу т.д. записуються без пробілу; скорочення та ін., і т.д., і т.п. можна використовувати лише в кінці речення, у середині їх слід записувати повністю.

Якщо при створенні документа не потрібно фіксувати увагу читача за допомогою рваної правої межі, то абзаци прийнято вирівнювати виключкою. Якщо просто (без переносів) вирівнювати текст за правою межею, виникає брак, який має назву рідкий рядок чи зяючі пробіли. Виключка надає тексту оформлений вигляд і має перевагу: вирівняний текст містить більшу кількість символів у кожному рядку, що зменшує загальну кількість рядків. Доцільно розставляти переноси слів на завершальній стадії редагування.

Існує дві можливості переносу: автоматичний – без контролю розробника документа та примусовий перенос із контролем виконання переносу користувачем. При підготовці важливих документів рекомендується використовувати примусовий перенос.

До основних операцій редагування відносять: додавання, знищення, переміщення, копіювання фрагменту тексту, а також операція пошуку і контекстної заміни.

Під *фрагментом* розуміється область тексту, вказана (виділена, маркована) користувачем. Мінімальний розмір фрагменту – один символ, максимальний – весь текст документа. Текстовий процесор Word Office 2000 позволяє зберігати до 12 фрагментів тексту одночасно і використовувати їх для редагування вибірково чи всіх одразу.

Параметри оформлення документа (приклад): символи – нормальної насиченості, кегль 10 пунктів; абзаци – без відступів, вирівняні вліво, через один інтервал; величина табуляції – через 0,5 дюйма (чи 1,27 см); розмір друкованої сторінки документа - формат А4 (210 мм на 297 мм); межі тексту на друкованій сторінці - ліве і праве 3,17 см, верхнє і нижнє 1,5 см.

Кожен документ створюється за якимось шаблоном уже існуючого документа.

Розрізняють три типи форматування прозаїчних документів: символьне (чи шрифтове оформлення); форматування абзацу документа; оформлення (верстка) сторінок (чи розділів) документа.

Шрифт – комплект літер із буквами для того чи іншого алфавіту й усіма знаками та цифрами, які до нього відносяться. Гарнітура – це графічні особливості шрифту, що визначаються нахилом (шрифт прямого чи курсивного накреслення), шириною (нормального, вузького і широкого накреслення), насиченістю (світлого, напівжирного, жирного накреслення).

Сімейство шрифтів – набір шрифтів, що мають багато спільного. Гарнітури поділяють на великі чотири категорії:

1. Serif (Сериф – шрифт із засічками). Найпопулярніші гарнітури Times, Bookman. Шрифт із засічками краще використовувати для основного тексту. При читання такого тексту погляд ніби ковзає вздовж лінії засічок.

2. Sans serif (Санс сериф – шрифт без засічок). Рублені шрифти. Одна з найпопулярніших гарнітур цієї категорії – Helvetica.

3. Script (Скрипт – імітація рукописного тексту).

4. Гарнітури типу «Рі» розробляються для спеціальних застосувань. Наприклад, для нот, формул хімічних сполук тощо. Windows використовує шрифт Wingding цієї гарнітури.

При виборі шрифту для друку документа основним критерієм є зручність читання. Обраний шрифт не повинен відволікати увагу читача від змісту документа. Різноманітність лише може внести текст заголовних надписів різних складових частин документа (рубрик). Для основного тексту рекомендується використовувати текст із засічками. Пряме призначення засічок – підвисити читабельність тексту. Рубрикація (система заголовків) повинна привертати увагу читача. Для такого тексту при виборі шрифту головний критерій – не розбірливість, а те, наскільки він контрастує з основним текстом. Рекомендується використовувати із цією метою рублений текст без засічок. Прикладом такого шрифту у Word для Windows може слугувати Arial, аналог шрифту Helvetica.

Загальні рекомендації зі створення сторінки документа: не використовуйте на одній сторінці багато видів гарнітур (максимально рекомендується три типи); всі сторінки документа набирайте обраним набором гарнітур; притримуйтесь сітки-схеми розміщення тексту та ілюстрацій на сторінці; не намагайтесь максимально заповнити весь простір сторінки текстом.

Стандартними параметрами оформлення символів є: тип (гарнітура) шрифту; кегль (висота) символів шрифту; накреслення літер (напівжирний, курсив, напівжирний курсив, звичайний); підкреслення; колір символів; розташування символів щодо опорної лінії рядка (верхній і нижній індекс). Оформлення абзаців документа (бесіда зі студентами).

Верстка сторінок багатосторінкового документа. Якщо система підготовки тексту використовується для створення і оформлення багатосторінкового документа, то в тексті можуть з’явитися нові структурні елементи: колонтитули, виноски, закладки, перехресні посилання. Під закладкою чи міткою розуміється певне місце в тексті документа, якому користувач присвоює ім’я. У подальшому закладка в багатосторінковому документі може використовуватись для: швидкого переходу до місця документа, позначеного закладкою; створення перехресних посилань у документі.

Іноді під час читання документа необхідні доповнення до основного тексту, підрядкові примітки, які оформляють виносками. Знак виноски розміщують в основному тексті біля того місця, якого стосується примітка, і на початку самої примітки. Рекомендується в текстовому матеріалі використовувати знак виноски у вигляді

арабських цифр, а в цифровому – у вигляді букв чи знаків. Перехресне посилання – це текст, що пропонує читачу документа звернутись до іншого фрагменту чи рисунку, що міститься в тексті. Наприклад: «Поверніться до розділу «Базові функції редагування тексту» (сторінка 145)».

Колонтитулом називається однаковий для групи сторінок текст (графічне зображення), розташоване поза основним текстом документа на полях друкованої частини. Розрізняють верхній (над тестом) і нижній (під текстом) колонтитул. Порядкові номера сторінок входять до колонтитула. їх називають колонцифрами. Стандартними параметрами оформлення сторінок документа є: поля сторінок, розмір друкованого аркуша і орієнтація тексту на папері; розташування колонтитулів; кількість колонок тексту (газетний стиль). Деякі програми класу СПТД дозволяють нумерувати текст документа не з першої сторінки, здійснювати розбиття на сторінки за правилами поліграфії. Наприклад, не допускаються так звані «висячі» рядки – одиночні рядки абзацу зверху чи знизу сторінки, може автоматично відстежуватись розташування двох сусідніх абзаців на одній сторінці документа (хоча б два рядки наступного абзацу розташовані на сторінці з попереднім абзацом тексту) тощо.

Сучасні інформаційні технології активно використовують ідею збереження не тільки конкретного результату роботи у вигляді документа, але і сукупності дій, за допомогою яких цей результат був досягнутим. Ця ідея в системах підготовки текстових документів реалізована у вигляді механізму шаблонів. Шаблон – це документ спеціального типу, який містить різноманітну інформацію про стилі форматування частин документа, вставлених полях тощо. Будучи один раз підготовленим і збереженим у пам'яті комп’ютера, шаблон дозволяє швидко виготовляти аналогічні за формою (але не за змістом) документи без затрат часу на форматування. Шаблон – це в загальному план тексту, графіки документа і набір способів форматування окремих його частин. У комплектах постачання текстових процесорів завжди містяться бібліотеки готових стилів, але готові шаблони багатьох СПТД не відповідають вимогам державного стандарту. При виготовленні власного шаблона бланку підприємства необхідно виконати певну послідовність кроків:

1. Встановити параметри сторінки бланка.

2. Помістити в бланк постійні реквізити згідно з вимогами державного стандарту чи міжнародних правил оформлення документів. При цьому використовуються всі можливості конкретної СПТД, наприклад вставка графічних зображень, емблеми підприємства, встановлення параметрів для шрифтового виконання різних частин майбутнього документа, встановлення атрибутів оформлення абзаців основної і заголовної частин документа тощо. Крім оформлення до шаблона може бути включений глосарій часто вживаних слів і фраз ділової лексики для даного типу документів.

3. Зберегти бланк як шаблон.

Створена таким чином колекція шаблонів документів використовується усіма співробітниками підприємства, що забезпечує уніфікацію оформлення документів підприємства і зменшує час на виготовлення конкретного документа.

Під «великим» документом у системі підготовки текстових документів розуміється документ, що має не тільки і не стільки великий обсяг, але й складну будову. До цього класу можна віднести різного роду статті, звіти, технічні описи, проектну документацію. Практично всі сучасні потужні текстові процесори мають широкий набір засобів роботи зі складними структурованими документами.

Під структурою документа розуміється схема, що визначає взаєморозташування і зв’язок його складних частин. Грамотне структурування документа підвищує його значимість і ступінь впливу на користувача. Для ефективної роботи з великими документами користувач текстового процесора має у своєму розпорядженні наступний набір операцій: створення структурованого документа і реорганізація його структури (підвищення чи пониження рівня ієрархії деяких заголовків); перегляд структури документа з виведенням на екран тільки заголовків певного рівня ієрархії; створення виносок, покажчиків, змісту, посилань, списку ілюстрацій, закладок.

Створення «великого» документа повинно починатись із розробки його структури, підпорядкування і ієрархії заголовків. Документ можна зробити привабливішим, якщо додати в нього графічні об’єкти. Після створення графічного об’єкта його можна залити кольором чи візерунком, змінити колір і тип ліній, збільшити чи зменшити, перемістити, повернути чи дзеркально

відобразити. При додаванні малюнка в документ він приєднується до тексту, коло якого знаходиться.

Більшість текстових процесорів підтримує концепцію складного документа – контейнера, що містить у собі об'єкти різних форматів. Користувач має можливість вставити і текст документа різні малюнки, таблиці, графічні зображення, підготовлені в інших програмних середовищах.

Розміщення графічних фрагментів у текстовому документі здійснюється з використанням кадрів. Кадр – сховище для розміщення об'єктів у частині сторінки, що не визначається параметрами полів друкованої сторінки (наприклад, між колонками тексту чи в частині полів сторінки). Одна з найважливіших властивостей кадру – можливість розміщувати довкола об'єкту текст. Розвинуті системи комп'ютерної підготовки текстів дозволяє використовувати кадри як прямокутної, так і неправильної форми. інша важлива властивість – здатність змінювати розмір і місце розташування на сторінці.

Більшість текстових процесорів забезпечує користувачу засоби написання макрокоманд. Макрокоманда – це опис на спеціальній мові часто використовуваних операцій, дій і фрагментів тексту. Основне призначення макрокоманд (чи макросів) – підвищення продуктивності праці користувача, автоматизація часто виконуваних команд.

Сьогодні текстові процесори все частіше входять до складу прикладних офісних програмних комплексів нового покоління. Крім пакету СПТД у комплекс входять електронна таблиця, програма презентаційної графіки і СУБД (система управління базами даних). Наприклад, текстовий процесор Word Perfect сьогодні не продається як автономний продукт, а тільки в складі пакету Corel Office Professional. На ринку прикладних програмних комплексів домінують три компанії: Microsoft, Corel і Lotus. Хоча кожному комплексу притаманні унікальні особливості, всі вони розвиваються за загальними законами. Передусім – це повна уніфікація: загальний інтерфейс і однакові підходи до вирішення таких типових задач, як управління файлами, редагування, форматування, друк, робота з електронною поштою і пошук підказки довідкової системи. У нових версіях програм класу СПТД велику увагу приділено інтеграції з глобальною мережею Internet. Ступінь їх сумісності зі старими версіями залишається високим.

Усі основні текстові процесори підтримують перевірку правопису «по ходу», мають багато можливостей настільних видавничих систем: «водяні знаки», буквиці, друк тексту в оборку графічної ілюстрації неправильної будови. Головною тенденцією у напрямі удосконалення засобів програмування стає міжпрограмна сумісність усіх продуктів комплексу.

Майже всі програми прикладних комплексів мають інструменти для колективної роботи: засоби редагування і збереження документів основних типів у форматі HTML для Internet. Сьогодні більша частина документів створюється в результаті колективної роботи багатьох людей. Звичайною практикою є коментування документа декількома людьми – і ці коментарі повинні бути знайдені і враховані. Розвинуті системи підготовки текстових документів надають засоби відмітки, відстежування і редагування виправлень, що значно спрощує процес обліку думок авторів рецензій. Для коментування документа без його зміни використовуються примітки – послідовно пронумеровані коментарі до різних місць документа. Зміни в текст вносяться в режимі запису виправлень. Записи виправлень показують, у яких місцях документа текст чи графіка були додані, знищені чи переміщені.

Переваги використання електронної пошти і списків розсилки загальновідомі: ці засоби дозволяють впорядкувати і надійним способом довести інформацію до всіх виконавців. Крім того, засіб сповіщення про вручення вбудований у більшість сучасних поштових систем, дозволяє контролювати час, коли інформація дійсно була одержана. Сучасні розвинуті текстові процесори мають убудовані системи електронної пошти і управління розсилкою, що дозволяє розробнику документа, не виходячи із середовища розробки, відіслати електронною поштою створений документ. Автори документів мають можливість визначити коло осіб, що мають доступ до документу, і права кожного з них щодо внесення змін: документ може бути призначеним тільки для читання, можна заборонити зберігати документ під будь-яким іншим іменем чи змінювати його стилі чи, навпаки, користувачу можуть надаватись повні права доступу. Новим засобом колективної роботи є функція відстежування версій і поява вікон з іменем автора і датою внесення змін.

Процес створення нових і удосконалення існуючих систем

текстової обробки триває. Ринок збуту програм цього класу є значним і має стійку тенденцію до подальшого розширення. Роль систем підготовки текстових документів буде постійно зростати в міру удосконалення інформаційних технологій у ході інформатизації суспільства.

Питання для самоконтролю

1. Які існують проблеми організації редакційно-видавничого процесу за допомогою комп'ютерних програм?
2. В чому особливості редакторської роботи з комп'ютерними засобами контролю правопису?
3. В чому особливості редакторської роботи при використанні автоматизованої системи оптичного розпізнавання тексту та його перекладу з іншої мови?
4. Які типові помилки автоматизованих програм перевірки правопису?
5. Які існують способи підготовки текстового матеріалу за допомогою комп'ютерів?
6. Назвіть основні правила комп'ютерного набору тексту.

Завдання

1. Опишіть програму Adobe PageMaker:
   1) Специфіка роботи програми верстки PageMaker.
   2) Що таке «текстовий блок»?
   3) Перерахуйте основні характеристики шрифту
   4) Які елементи належать до графічних?
   5) Що таке «символи, що не друкуються»? Для чого вони використовуються?
   6) Як створити колонки для текстового блоку?
   7) Яким чином можна розміщувати текст у текстовий блок? Перерахуйте та опишіть всі види розміщення.
   8) Як можна здійснювати зміну текстового блоку?
   9) Що так абзац із точки зору програми PageMaker?
   10) Які абзац має основні характеристики?
   11) Що таке «стиль» і для чого він використовується?
   12) Як можна вставляти / видаляти сторінки?
   13) Чому не рекомендується здійснювати редагування тексту в програмі верстки?
   14) Як вносити зміни в текст у текстовому режимі?
   15) Як знайти / замінити в тексті необхідну фразу?

16) Для чого призначений текстовий редактор?
17) Які види інтервалів ви знаєте? Що таке «кернінг»?
18) Як може здійснюватися перенос слів у рядках?
19) Які види ініціалів ви знаєте? Яким чином можна розміщувати ініціали у тексті?
20) Як розташувати текст таким чином, щоб окремі його частини в рядках розміщувалися, починаючи з визначеної позиції?
21) Які способи оформлення нумерованих та маркованих списків використовуються у програмі?
22) Для чого використовується знак табуляції?
23) Як можна відформатувати текст у комірці та комірку в цілому?
24) Що таке «фрейм»?
25) Як створити фрейм?
26) Яким чином можна заповнити фрейми?
27) Опишіть особливості роботи зі зв'язаними текстовими фреймами.
28) Як можна сортувати сторінки у публікації.
29) Яким чином можна організувати складну публікацію?
30) Як можна створити зміст?
31) Як здійснюється попередній запис списку тем?
32) Які основні характеристики предметного покажчика?
33) Яким чином можна розтягнути / ущільнити текст.

2. Опишіть програму Adobe InDesign:

1) Специфіка використання палітр у програмі InDesign.
2) Варіанти створення нового документу.
3) Що таке «текстовий блок»?
4) Які елементи належать до графічних?
5) Що таке «символи, що не друкуються»? Для чого вони використовуються?
6) Як створити горизонтальні та вертикальні колонки для текстового блоку?
7) Як зафіксувати колонки? Як змінити розташування на сторінці колонок?
8) Яким чином можна розміщувати текст у текстовий блок? Перерахуйте та опишіть всі види розміщення.
9) Як можна здійснювати зміну текстового блоку?
10) Які абзац має основні характеристики?

11) Чому не рекомендується здійснювати редагування тексту в програмі верстки?
12) Як вносити зміни в текст у текстовому режимі? Як знайти / замінити в тексті необхідну фразу?
13) Для чого призначений текстовий редактор?
14) Які види інтервалів ви знаєте?
15) Як може здійснюватися перенос слів у рядках?
16) Яким чином можна розміщувати буквиці?
17) Як розташувати текст таким чином, щоб окремі його частини в рядках розміщувалися, починаючи з визначеної позиції?
18) Для чого використовується знак табуляції?
19) Як можна відформатувати текст у комірці та комірку в цілому?
20) Як замінити текстовий матеріал на зверстаній сторінці?

МЕТОДИЧНІ МАТЕРІАЛИ ДЛЯ САМОСТІЙНОЇ РОБОТИ

Самостійна робота – це форма організації індивідуального вивчення студентами навчального матеріалу в аудиторний та позааудиторний час.

Самостійна робота з курсу «Прикладне мовознавство (Квантитативна і комп'ютерна лінгвістика, в т.ч. автоматична обробка природної мови)» передбачає пошук та опрацювання рекомендованої літератури, виконання завдань, спрямованих на розвиток самостійності та ініціативності.

Для самостійної роботи з «Прикладного мовознавства (Квантитативна і комп'ютерна лінгвістика, в т.ч. автоматична обробка природної мови)» пропонуються завдання, покликані розширювати, поглиблювати, закріплювати базові знання, отримані під час лекційних (пропонуються теми концептуального характеру) та практичних занять.

Для контролю знань студентів використовуються:

а) письмові відповіді на теоретичні питання;

б) письмові роботи;

в) практичні завдання із комп'ютером.

Виконання самостійних завдань є обов'язковою умовою допуску до підсумкової контрольної роботи. Матеріали самостійних завдань подавати на перевірку викладачу в окремому зошиті (або в електронному варіанті відповідно до характеру завдань).

1. Створення мультимедійних презентацій.

Мультимедійна презентація – це сукупність текстів, зображень, звуку, анімації та інших засобів представлення інформації на визначену тему, яка зберігається у файлі спеціального формату з розширенням Ppt. Їх використання дозволяє досягти максимальної ефективності презентації інформації, забезпечуючи одночасний вплив на зорові і слухові органи чуття слухачів. Для створення мультимедійних презентацій використовуються різноманітні програми, але найбільш доступним засобом для отримання власних комп'ютерних навчальних продуктів є програма PowerPoint – майстер створення презентацій, яка входить до складу інтегрованої системи Microsoft Offise. Мультимедійні презентації, створені у PowerPoint, дозволяють усвідомлено і гармонійно інтегрувати багато видів інформації.

Навчальна інформація може представляться в різних формах:

1) зображення, включаючи фотографії, малюнки, карти, високоякісну графіку тощо;

2) звук, у тому числі, і стерео: звукозаписи голосу, звукові ефекти і музику;

3) відео, відеоефекти, рухоме відеозображення;

4) анімації й анімаційні імітування.

Етапи підготовки мультимедійної навчальної презентації.

Етап планування: визначення призначення презентації і цільової аудиторії.

Етап проектування: складання сценарію реалізації інформації; визначення змісту кожного слайду і їх послідовності; розробка дизайну.

Етап інформаційного наповнення: підготовка медіафрагментів: структурування і відбір тексту навчального матеріалу для слайдів, ілюстрацій – сканування малюнків, відео, запис аудіофрагментів; підготовка мовного і відеосупроводу (у разі необхідності).

Етап створення: наповнення слайдів медіафрагментами, тобто підготовленим текстовим та ілюстративним матеріалом; створення дизайну слайдів.

Етап налаштування: настроювання анімаційних ефектів, керуючих кнопок; встановлення гіперпосилань на елементи для виходу в Інтернет і поєднання зовнішніх програм; окремо робиться музичний супровід; слайди записуються в пам’ять комп’ютера.

Етап тестування: наприкінці проводиться тест перевірка готової презентації, а саме – виправлення помилок у тексті й ілюстраціях; узгодження анімаційних ефектів; перевірка гіперпосилань.

Етап друкування: не є обов’язковим, але коли матеріал презентації необхідно розповсюдити серед слухачів, тоді виконують друкування слайдів.

Етап використання: виступ на заняттях, демонстрація на конференціях, виставках тощо.

Етап удосконалення: внесення змін до сценарію, навігаційної схеми, матеріалу, що складає змістову частину презентації чи її ілюстративне доповнення.

Під час створення мультимедійної презентації необхідно враховувати не тільки відповідні принципи класичної дидактики,

але й специфічні підходи використання комп'ютерних мультимедійних презентацій. Фахівцями на сьогодні вироблені певні вимоги до підготовки презентацій. Їх дизайн повинен бути простим, але ефективним. Презентація слугує лише фоном для усного повідомлення матеріалу, вона не повинна відволікати увагу від лектора, аудиторія має слухати й сприймати матеріал, а не лише переглядати картинки на екрані. Задля цього всі слайди навчальної презентації слід оформлювати в єдиному стилі, який при цьому не відволікатиме від самої презентації, а керівні кнопки не робити яскравішими за основний зміст слайда. Всі слайди повинні витримуватись в єдиному стилі. Часто верхня і нижня частини слайду погано відображаються на екрані, тому не потрібно розміщувати там важливу інформацію. Необхідно визначити стиль – шрифт і спосіб представлення тієї інформації, яка буде повторюватись від слайду до слайду. Найбільш важлива інформація – висновки, визначення, правила тощо, представляється крупним і виділеним шрифтом 24 розміру. Головний текст повинен, бути, як мінімум, 18 розміру. Доцільно використовувати різні прийоми для виділення найбільш важливої інформації; слайди не повинні бути занадто яскравими, зайві прикраси лише створюють бар'єр на шляху ефективного передавання інформації; інформація на слайді повинна бути добре структурована. Необхідно враховувати, що люди можуть одночасно запам'ятовувати не більше трьох фактів, висновків, визначень. Дієслова повинні вживатися в одній часовій формі, заголовки приваблювати увагу аудиторії й узагальнювати ключові положення слайду. Використовуючи статистичні дані для обґрунтування своїх ідей, не потрібно розміщувати на одному слайді більше чотирьох блоків інформації. Крім того, під час оформлення презентації не доцільно використовувати: надто яскраві кольори (рекомендують застосовувати для фона холодні тони зелено-синьої гами і невиразні текстури); більше, як три кольори на одному слайді (при цьому кольори тексту та фону мають бути контрастними; рекомендують застосування кольорових схем «світлий текст на темному фоні» і «темний текст на білому фоні»); значну кількість анімаційних ефектів (їх застосування має бути обґрунтованим); довгі речення і формулювання; невиразні й непомітні заголовки; завелику чи замалу кількість слайдів. Найбільш оптимальною є презентація із 36 слайдів для 10 хвилин – таке співвідношення можна використовувати у випадку, якщо

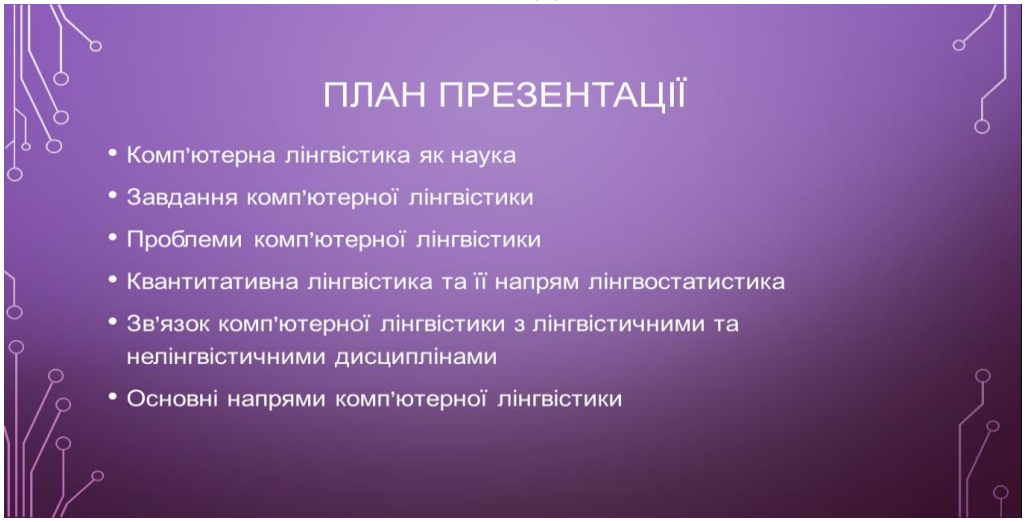
слайд може бути повністю сприйнятий свідомістю аудиторії за 15 секунд.

Теми для створення презентацій

1. Комп’ютерний морфологічний аналіз
2. Автоматичний синтаксичний аналіз
3. Автоматичний семантичний аналіз
4. Бази даних і бази знань
5. Семантичні мережі й автоматичне опрацювання тексту
6. Автоматичне реферування й анотування тексту
7. Машинний переклад як різновид інтелектуальних систем АОТ.

Зразки слайдів

Слайд 1



Слайд 2

2. Створення електронних словників.

Власне електронний (мультимедійний) словник – це словник, укладений на основі фактичного матеріалу і створений за допомогою комп'ютерних програм.

Велике значення для сприйняття навчального контенту має форма його подання. Вибір шрифтів, палітри кольору для фону і тексту, вирівнювання і вибір міжрядкових інтервалів та ін, тобто форматування тексту, а також графічне оформлення є важливим етапом роботи з навчальною інформацією. Слід також використовувати різноманітні прийоми, що враховують вікові психолого-фізичні особливості сприйняття людиною текстової інформації з екрана монітора, які могли б істотно поліпшити дизайн-ергономіку тексту.

Важливим із принципів створення електронних словників є принцип реалізації структури гіпертексту. Навчальні тексти повинні представляти собою особливим чином організований багаторівневий гіпертекст, що дозволяє здійснювати вивчення предмета з встановленням різних логіко-семантичних відносин, компенсувати дидактичні втрати внаслідок відсутності аудиторної навчального середовища (реалізація переходів на різні додаткові, інформаційні та графічні об'єкти за посиланнями).

Гіпертекст – одна з найважливіших характеристик електронних навчальних видань. Гіперпосилання в Microsoft PowerPoint дає змогу змінити порядок переходу з одного слайда на інший, відкривати веб-сторінки або інші файли за допомогою відповідних програм. Словник містить гіперпосилання, а тому має розгалужену структуру, оскільки в режимі демонстрації користувач може обрати за власним бажанням послідовність відображення слайдів за допомогою запропонованих гіперпосилань. Гіперпосилання можна додати як до тексту, так і до малюнка чи іншого об'єкта слайда. Виконати перехід за гіперпосиланням можна лише в режимі демонстрації. Фрагмент тексту, що є гіперпосиланням, відрізняється від іншого тексту на слайді: крім підкреслювання, цей фрагмент має інший колір символів, що визначається колірною схемою обраного шаблону оформлення. В режимі показу слайдів при наведенні вказівника мишки на гіперпосилання він набуває вигляду руки.

Щоб додати гіперпосилання, необхідно виділити на слайді потрібний об'єкт, наприклад фрагмент тексту, зображення тощо, та виконати вказівку меню Вставка / Гіперпосилання або відповідну

вказівку контекстного меню.

У лівій частині діалогового вікна Додавання гіперпосилання потрібно вибрати тип об’єкта, на який буде здійснюватися перехід за гіперпосиланням: файл або веб-сторінка, місце в документі (тобто інший слайд поточної презентації), новий документ чи електронна пошта.

Кожен словник має вміщувати не менше 10 термінів.

Теми для створення електронних словників

1. Прикладні аспекти морфології, лексикографії та синтаксису
2. Аналітико-синтетичне опрацювання документів
3. Лінгвістичне моделювання й комп’ютерні лінгвістичні системи
4. Проблеми створення систем штучного інтелекту

3. Створення елементів відеолекцій на теми.

Відеолекція – мультимедійна оффлайн презентація, де на основному екрані зазвичай відображається ілюстративний матеріал лекції – слайди, що містять текст і графіку, в кутку екрану можна бачити відео із зображенням лектора або інші відеоматеріали, а внизу розміщені засоби навігації, які дозволяють зупиняти і поновлювати перегляд відеолекції, перемотувати її вперед і назад, переходити до бажаного розділу тощо.

Програми для створення відео лекцій: iSpring Pro, Camtasia Studio, Модуль Rich Media та інші.

Модуль Rich Media– це перший безкоштовний засіб створення відеолекцій, який до того ж надає звіти про те, хто коли і скільки раз слухав відеолекцію. Rich Media – це засіб створення і відображення мультимедійних презентацій (відео + слайди синхронізовано) в курсі Moodle. Він був розроблений французькою компанією Symetrix в рамках проекту з електронного навчання для CNEMI.

Елемент відеолекції передбачає п’ятихвилинний виступ студента, що стисло, доступно, повно розкриває актуалізовану проблему.

Теми для створення відеолекцій

1. Історія розвитку комп’ютерної лінгвістики.
2. Спеціальні системи письма

4. Конспект матеріалу за навчальною літературою.

Конспектування є процесом розумового переосмислення і письмової фіксації прочитаного тексту. Внаслідок конспектування з'являється запис, який допомагає його автору негайно чи через деякий час із необхідною швидкістю відтворити отриману раніше інформацію. Конспектування дозволяє вичерпно викласти головний зміст творів, документів, з'ясувати внутрішній зв'язок і логічну послідовність обґрунтованих у них теоретичних положень.

До конспектування слід приступати лише після загального ознайомлення з його змістом, засвоєння зв'язку між основними думками, положеннями, головною ідеєю твору. Конспекти, складені без попереднього прочитання першоджерела, перенасичені другорядними відомостями. При цьому необхідно пам'ятати головні вимоги до конспектування. Конспектування є творчим процесом. За змістом і формою конспект повинен бути складений так, щоб допомагати засвоєнню головних положень праці в їх логічній послідовності, швидкому, глибокому запам'ятовуванню і відтворенню прочитаного. Важливою вимогою до конспектування і головною його перевагою є висловлювання студентом свого ставлення до прочитаного.

Дбаючи про раціональну організацію тексту конспекту, слід дотримуватися таких правил:

– чітко, стисло, лаконічно формулювати думку. Це дає можливість зосередитися на головному, найсуттєвішому в прочитаному тексті;

– дослівно занотовувати визначення, афористичні думки, аргументи автора. Думку, цитату не можна обривати посередині, за необхідності зробити пропуски в тексті використовують три крапки;

– всі цитати слід брати в лапки, точно зазначати джерело (назва, місце виходу у світ, видавництво, рік випуску, том, сторінка);

– оформлювати важливі статистичні дані у вигляді таблиць, графіків, діаграм;

– використовувати скорочення слів, умовних позначень, якщо вони цілком зрозумілі автору конспекту і не ускладнюють перечитування записів;

– записувати компактно, що дозволяє зробити конспект доступним для огляду;

– робити між рядками інтервали, достатні для вписування доповнень за необхідності;

– датувати записи.

Окремі доповнення можна записувати на аркушах чи картках, які вкладають у конспект, за необхідності використовуючи їх. Обсяг конспекту залежить від рівня характеру, складності та обсягу джерела. Багатослівні, великі за обсягом конспекти, можуть свідчити про недостатню осмисленість роботи студента. Виконаний із дотриманням головних вимог конспект сприяє засвоєнню знань, є допомогою в підготовці до іспитів, у майбутній практичній діяльності.

Питання для конспектування

1. Збереження, перетворення і захист документів. Відкривання текстових файлів. Режими перегляду сторінок. Друкування документа.
2. Копіювання файлів. Створення резервної копії файлу. Нумерація сторінок. Колонтитули. Оформлення посилань.
3. Уведення і виділення тексту: Видалення, заміна й переміщення тексту,його фрагментів. Підкреслювання і виділення кольором тексту. Вставка символів і знаків. Маркери у текстовому документі. Робота з таблицями, фігурами, малюнками, діаграмами.
4. Форматування документа. Команда «Абзац» (визначення відступів, міжрядкового інтервалу, способу вирівнювання та положення тексту на сторінці). Вибір шрифтів, їх види – Arial, Times New Romans, Calibri та ін. (звичайний, курсив, напівжирний, напівжирний курсив) та розміри.
5. Розмітка сторінки, вибір її параметрів (поля, масштаб, розмір паперу, орієнтація – книжкова чи альбомна).
6. Послідовність процесу комп’ютерного редагування: етапи тексту.
7. Технологічні особливості комп’ютерного редагування. Його відмінність від традиційного редагування.
8. Характеристика стандартних текстових програм WordPad, Блокнот.
9. Переваги нової версії програми Microsoft Word.
10. Застосування комп’ютерних систем Плай і Рута для перекладу й редагування тексту.
11. Довідкові ресурси Word.
12. Тенденції інформатизації українського суспільства.

13. Розвиток інформаційно-соціальних технологій в редакціях ЗМІ.

14. Незручності. пов’язані з використанням професійного програмного забезпеченя журналістами та редакторами.

15. Позитивні сторони процесу комп’ютеризації українських масмедіа.

16. Комп’ютерне редагування як творчий процес.

17. Комп’ютерні системи – причина зміни технології редакторської праці. Феномен електронного «пера». Проблеми організації редакційно-видавничого процесу за допомогою комп’ютерних програм.

18. Особливості редакторської роботи з комп’ютерними засобами контролю правопису.

19. Межі і системи атоматизації редакційно-видавничої діяльності.

20. Підходи сучасних науковців до оцінки ролі коректора у процесі комп’ютерного редагування.

21. Особливості редакторської роботи при використанні автоматизованої системи оптичного розпізнавання тексту та його перекладу з іншої мови.

22. Типові помилки автоматизованих програм перевірки правопису.

23. Характеристика комп’ютерної програми: функції, можливості.

24. Поняття тексту, його визначення. Історія виникнення тексту. Текст як основний матеріал друкованих і електронних видань.

25. Способи підготовки текстового матеріалу за допомогою комп’ютерів.

26. Основи і правила комп’ютерного набору тексту.

27. Вимоги до оформлення наукових статей.

28. Структурні елементи тексту (абзац, параграф, рубрика, розділ, підрозділ, шапка, заголовок, підзаголовок, пункти, підпункти та ін.).

Література

1. Городенко Л. Системи верстки: Практ. посіб. для студ. / Л. Городенко – К. : Центр вільної преси, 2007. – 520 с.

2. Іванов В. Ф. Комп’ютерні мас-медіа на межі століть // Актуальні питання масової комунікації. – 2002. – Вип. 3 (Ч. 1). –

С. 41-43.

3. Карпенко В. О. Основи редакторської майстерності. Теорія, методика, практика: Підручник. / В. О. Карпенко – К. : Університет «Україна», 2007. – 431 с.

4. Комолова Н. В. Компьютерная верстка и дизайн / Н. В. Комолова. – Спб : БХВ-Петербург, 2003. – 317 с.

5. Партико З. В. Галузеве редагування в засобах масової інформації: Конспект лекцій. / З. В. Партико – Львів : Афіша, 2007. – 104 с.

6. Партико З. В. Загальне редагування: нормативні основи: Навч. посіб. / З. В. Партико – Львів : Афіша, 2006. – 416 с.

5. Виконання практичних завдань.

І. Статистичний аналіз

Завдання: використовуючи дві із чотирьох запропонованих програм статистичного аналізу, здійсніть статистичний аналізу текстів, порівняйте отримані результати та поясніть вибір програм.

Програми:

http://www.textanalyzer.ru/,
http://advego.ru/text/seo/top/,
http://www.practiline.com/download.htm,
http://www.blog-kaplunoff.ru/programmy-dlya-kopirajterov.html

Рекомендації щодо використання програми Textus Pro

Програма Textus Pro справжня знахідка для копірайтерів, вебмастерів, оптимізаторів. Цей простий, інтуїтивно зрозумілий та разом із тим багатофункціональний інструмент допоможе створити або оптимізувати текст, враховуючи алгоритми роботи пошукових систем. Тобто після аналізу тексту за допомогою програми Textus Pro та внесення необхідних корективів, матеріал стане більш значущим для пошукачів.

Рекомендації щодо використання програми PractiCount and Invoice

PractiCount and Invoice – програма для підрахунку слів, символів та створення інвойсів для перекладачів, котра буде корисною для тих, хто хоче отримувати реальну статистику практично з усіх видів офісних документів. Програма рахує слова, рядки, символи, сторінки та видає іншу статистику по тексту, котра налаштовується користувачем.

За допомогою цієї програми ви можете порахувати: слова,

символи з пробілами, символи без пробілів, рядки, умовні рядки, сторінки, друковані сторінки, повтори слів, японські, китайські та корейські слова та символи**.** PractiCount може рахувати та аналізувати текст у таких форматах: Microsoft Word (doc, docx, rtf), Microsoft Excel (xls, xlsx, csv), Microsoft PowerPoint (ppt, pptx, pps), Corel Word Perfect (wpd), Adobe Acrobat (pdf), Adobe Framemaker (mif), HTML (htm, html, shtml), XML, SGML, ASP, PHP, Help files (cnt, hhc, hpj, hhk, hhp), OpenOffice (odp, ods, odp).

Рекомендації щодо використання програми ADVEGO

Цей сервіс підтримує більше ніж 80 мов, пропонує статистику та частоту окремих слів, стоп-фраз та фраз. Також є можливість перевірити текст на унікальність та орфографію.

Рекомендації щодо використання програми TextAnalyzer

http://www.textanalyzer.ru/

Цей сервіс пропонує детальну статистику тексту. Окрім загальних кількісних характеристик можна встановити також середнє значення, водність тексту, актуальні послідовності слів та частотний словник.

ІІ. Синтаксичний аналіз.

http://nlp.stanford.edu:8080/parser/index.jsp

Рекомендації щодо використання програми Stanford Online Parser

Багатомовний онлайн парсер, що підтримує три мови (арабську, китайську, англійську та іспанську). Із переваг можна зазначити високу швидкість обробки, детальний розбір, зручність застосування.

https://www.connexor.com/nlplib/?q=demo/syntax

Рекомендації щодо використання програми Machinese Syntax Demo

Демонстративна версія мультимовного парсера (фінська, англійська, іспанська, французька, шведська, німецька) з можливістю вибору типу представлення аналізу. Містить список з описами тегів. Може не працювати з деякими браузерами (Firefox), низька швидкість обробки.

http://www.link.cs.cmu.edu/link/submit-sentence-4.html

Рекомендації щодо використання програми Link Grammar

Англомовний парсер із простим оформленням та мінімальною кількістю функцій. Вихідні речення для розбору можуть уводитися вручну з клавіатури або задаватися в ASCII-файлі для пакетної

обробки. Результати розбору представляються у вигляді дерев лінгвістичного розбору (графи), що послідовно видаються програмою. Дерева лінгвістичного розбору впорядковані за імовірністю правильного розбору. Семантичні словники не використовуються. Має консольний інтерфейс. Вихідні коди та програма розповсюджуються безкоштовно.

http://slashzone.ru/parser/

Рекомендації щодо використання програми Граматика зв'язків

On-line програма синтаксичного аналізу речень російської мови. Створена на основі образу Link Grammar Parser. Алгоритм роботи синтаксичного аналізатора ґрунтується на використанні розробленої граматики зв'язків для російської мови. Доступний для тестування web інтерфейс програми. Морфологічний словник використовується від aot.ru. Програма реалізована на на unix C, Perl під ліцензією Apache License. Програма та вихідні коди розповсюджуються на комерційній основі.

http://mytts.forum2x2.ru/t418-topic

Рекомендації щодо використання програми Cognitive dwarf

Програмний пакет вміщує синтаксичний аналізатор для російської та англійської мов та систему автоматичного перекладу (російсько-англійськийта англо-російський).

III. Автореферування

Завдання: використовуючи запропоновані сервіси, виконайте автореферування текстів різних стилів. Проаналізуйте отримані результати. Автореферат тексту якого стилю виконано найкраще? Чому? Який із сервісів виконує автореферування краще?

Параметр аналізу	**Коментар студента**
Зв'язний текст або набір словосполучень / речень	
Функціональна завантаженість елементів реферату	
Чи відображені необхідні структурні компоненти реферату (тема мета)	
Загальний висновок	

http://visualworld.ru/referat.jsp

Сервіс реферування призначений для автоматичного

створення рефератів (анотацій) до текстів природною мовою. Сервіс виокремлює основні речення із тексту. Для отримання реферату скопіюйте текст та вставте його у вікно «Вихідний текст» та натисніть клавішу «Створити автореферат!». Після обробки в лівому стовбці буде показано вихідний текст, а у правому – отриманий реферат. Натискаючи клавіші «Зменшити» або «Збільшити» можна керувати рівнем стиснення тексту (розміром реферату). Якщо реферат вийшов пустим, то це, як правило, означає, що вихідний текст надто короткий та з нього не вдалося виокремити основні смислові речення.

http://textcompactor.com/

Цей безкоштовний онлайн інструмент створено, щоб допомогти читачам обробляти велику кількість інформації. Після того як ви завантажили свій текст, веб-сервіс вираховує частоту кожного слова в уривкові. Потім здійснюється підрахунок частоти кожного речення, пов'язаного зі словами, котрі воно містить. Автореферування краще за все здійснюється для підручників та довідкових матеріалів. Результати можуть бути спотворені, якщо текст містить лише декілька фраз. Text Compactor не рекомендується для автореферування текстів із вимислом.

IV. Морфологічний аналіз

Завдання: за допомогою сервісу http://starling.rinet.ru/morph.htm виконайте морфологічний аналіз слів: «кислород», «обезвредил», «научно-исследовательский», «expert», «stimulate» и «favourite». Які можливості пропонує сервіс для російської та англійської мови? Чи зустрічаються помилки? Поясніть причини.

http://starling.rinet.ru/morph.htm

Програма морфологічного аналізу слів російської та англійської мов. Працює із мультимовними текстами, зі знаками транскрипції. Реалізовано пошук, аналіз та синтез словоформ за словником Залізняка, з перекладом за словником Мюллера.

ПРИКЛАДИ МОДУЛЬНИХ КОНТРОЛЬНИХ РОБІТ

Модульна контрольна робота № 1

І варіант

І рівень

1. Наука, що вивчає системи керування:
   А) Прикладна лінгвістика
   Б) Комп’ютерна лінгвістика
   В) Кібернетика
   Г) Корпусна лінгвістика
2. Хто автор класифікації типів інформації?
   А) Партико З.
   Б) Дарчук Н.
   В) Білецький А.
   Г) Волошин В.
3. У скількох способах свого існування виявляється мова?
   А) 2
   Б) 3
   В) 4
   Г) 5
4. Наука про типи інформації:
   А) Інформатика
   Б) Кібернетика
   В) Корпусна лінгвістика
   Г) Комп’ютерна лінгвістика
5. Скільки операцій виконували комп’ютери першого покоління?
   А) 1000
   Б) 5000
   В) 10000
   Г) 100000
6. Яка елементарна база була у комп’ютерів третього покоління?
   А) Лампи
   Б) Інтегральні схеми
   В) Транзистори
   Г) Мережа транзисторів
7. Коли укладено шрифти UNICODE?
   А) 50-ті роки ХХ ст.
   Б) 70-ті роки ХХ ст.
   В) 90-ті роки ХХ ст.
   Г) п.ХХІ ст.

8. Яка одиниця виміру тривалості звука?
   А) Децибели
   Б) Мілісекунди
   В) Герци
   Г) Секунди
9. Для визначення тривалості звука використовують:
   А) Осцилографію
   Б) Спектрографію
   В) Палатографію
   Г) Рентгенографію
10. Які частоти не сприймає слух людини?
   А) Вищі за 20 кГц
   Б) Нижчі за 20 кГц
   В) Вищі за 20 Гц
   Г) Вищі за 100 Гц
11. Не є різновидом письма:
   Піктографічне
   Ідеографічне
   Силабічне
   Силабо-тонічне
12. Скільки принципів лежить в основі орфографії?
   А) 3
   Б) 4
   В) 5
   Г) 6
13. У формулі H=G/S, G – це:
   А) Кількість букв
   Б) Кількість звуків
   В) Кількість графем
   Г) Надлишковість абетки
14. Скільки точок для кодування використовує брайлівська система?
   А) 4
   Б) 5
   В) 6
   Г) 7
15. Утилітарна функція інформації реалізується шляхом :
   А) Зберігання інформації
   Б) Перетворення інформації

В) Передавання інформації
Г) Використання інформації

ІІ рівень

Дайте визначення термінам: надлишковість абетки, комп'ютерна лінгвістика, інформація, лігатура, стенографія.

ІІІ рівень

1. Розкрийте зв'язки комп'ютерної лінгвістики з іншими науками.
2. Охарактеризуйте завдання комп'ютерної лінгвістики в галузі дослідження спеціальних систем письма.

ІІ варіант

І рівень

1. З якою дисципліною КЛ пов'язана системою автоматичного перероблення тексту?
   А) Лінгвістика тексту
   Б) Автоматична обробка природної мови
   В) Кібернетика
   Г) Інформатика
2. За способом оформлення інформація буває:
   А) Логічна/естетична
   Б) Узуальна/спеціальна
   В) Вербальна/екстра вербальна
   Г) Кодована/некодована
3. Скільки функцій інформації виділив Білецький А.?
   А) 7
   Б) 6
   В) 5
   Г) 3
4. Що не є об'єктом дослідження КЛ?
   А) Продукт мовлення
   Б) Мовлення
   В) Мовна система
   Г) Мовна діяльність
5. Скільки операцій виконували комп'ютери третього покоління?
   А) 10000
   Б) 50000
   В) 100000

Г) 1000000

6. Коли почала формуватися КЛ?
   А) 50-ті роки XX ст.
   Б) 80-90 ті роки XX ст.
   В) 60-70 ті роки XX ст.
   Г) Кінець 90-х XX ст.
7. Чим ознаменувалися 40-50-ті роки XX століття?
   А) Створено перші електронні словники
   Б) Здійснено синтаксичний аналізу тексту
   В) Виникли комп'ютери
   Г) Виникла фірма Apple
8. У чому вимірюється частота звука?
   А) Герцах
   Б) Мілісекундах
   В) Децибелах
   Г) Секундах
9. Для вимірювання сили артикуляції застосовують:

Палатографію
Тензопалатографію
Спектрографію
Осцилографію

10. Що не відноситься до фізичної характеристики звука?
   А) Висота
   Б) Частота
   В) Інтенсивність
   Г) Тривалість
11. Ієрогліфічне письмо це:
   А) Піктографічне
   Б) Літературно-звукове
   В) Ідеографічне
   Г) Складове
12. Який компонент не становить систему письма?
   А) Напрям письма
   Б) Орфоепія
   В) Орфографія
   Г) Графіка
13. У формулі A=L/S, S – це:
   А) Кількість букв
   Б) Кількість графем

В) Кількість звуків

Г) Оптимальність абетки

14. Який спосіб кодування традиційного письма застосовують в електронних каналах звука?

А) Шифрування

Б) Азбука Морзе

В) Криптографування

Г) Шрифт Брайля

15. Дименсійна функція інформації реалізується шляхом :

А) Зберігання інформації

Б) Перетворення інформації

В) Передавання інформації

Г) Вимірювання інформації

ІІ рівень

Дайте визначення термінам: кібернетика, оптимальність абетки, криптографування, ієрогліф, графема.

ІІІ рівень

1. Проаналізуйте основні шляхи розвитку наукових досліджень у сфері комп'ютерної лінгвістики.

2. Охарактеризуйте завдання комп'ютерної лінгвістики в галузі прикладної фонетики.

Модульна контрольна робота № 2

І варіант

І рівень

1. Основою аналітико-синтетичного опрацювання документів є:

А) згортання інформації

Б) розгортання інформації

В) спрощення інформації

Г) ускладнення інформації

2. Процес пошуку за ознаками, зазначеними у запиті:

А) адресний

Б) семантичний

В) документальний

Г) фактографічний

3. За рівнем узагальнення інформації та впливом на аудиторію виокремлюють такі групи оглядів:

А) інформаційні та аналітичні

Б) цілісні та ділені

В) універсальні та тематичні

Г) огляди дійсності та огляди творів

4. Аналітико-синтетичне опрацювання документів – це

А) укладання вправ для відповідних рівнів вивчення мови;

Б) аналіз документів і наявної в них інформації, який дає змогу синтезувати за певними правилами інший текст;

В) створення текстів із застосуванням зовнішньої інформації;

Г) добір текстів, що відповідають встановленим морфологічним, лексичним і синтаксичним вимогам.

5. В середньому до 600 знаків містить

А) реферування

Б) редагування

В) коректура

Г) анотування.

6. Який вид аналітико-синтетичного опрацювання текстів через його найвищу складність практично неможливо автоматизувати?

А) індексування

Б) переклад

В) готування оглядів

Г) реферування.

7. Однією з перших і найвідоміших машин, що продемонструвала вміння вести діалог була

А) система «Еліза»

Б) система «Елочка»

В) машина Тьюрінга

Г) машина «Енігма».

8. 7 червня 2014 року відбувся конкурс, на якому

А) була представлена машина Тьюрінга;

Б) компёютерна програма Євген Густман, що видавала себе за 13-річного хлопчика з України, переконала 33% суддів, що вона людина;

В) Ерл Хант видав монографію «Штучний інтелект»;

Г) працівники Київського національного університету імені Тараса Шевченка створили навчальну програму спілкування з комп’ютером «Елочка».

9. На основі попередньо укладених словників ключових слів і словосполучень, анти ознак, списків реляторів, фреймів базується:

А) фреймове реферування
Б) фреймове анотування
В) тезаурусне реферування
Г) тезаурусне анотування.
10. Два основні різновиди рефератів:
А) монографічні, звітні
Б) фрагментні, аспектні
В) продуктивні, репродуктивні
Г) конспекти, резюме.

ІІ Рівень

1. Алгоритм діалогу з автоматом для моделювання мисленнєвих процесів, реалізованих у діалозі між людьми – це ….
2. Група взаємопов'язаних за змістом і за допомогою синтаксичних засобів речень, які порівняно з окремими реченнями виражають розвиток думки…

ІІІ рівень

1. Розкрийте основні принципи індексування.
2. Проаналізуйте основні критерії якості інформаційного пошуку.
3. На яких завданнях базуються системи комп'ютерного редагування.

ІІ варіант

І рівень

1. Вид аналітико-синтетичного опрацювання документів, що передбачає ідентифікацію змісту документа з метою його подальшого пошуку:
А) реферування
Б) атрибуція
В) індексування
Г) інформаційний пошук
2. Вдумливе, інтенсивне читання, під час якого відбувається запам'ятовування змістової інформації тексту:
А) ознайомлювальне
Б) реферативне
В) контрольне
Г) вивчаюче
3. До позамовних параметрів атрибуції текстів відносять:
А) знаки пунктуації

Б) структурні дані

В) помилки

Г) морфологічні параметри

4. Вид аналітико-синтетичного опрацювання текстів документів, у процесі якого генерують короткий текст (в середньому до 1000 знаків), що передає основний зміст документа називається

А) індексуванням

Б) реферуванням

В) анотуванням

Г) інформаційним пошуком.

5. Серед норм якого виду аналітико-синтетичного опрацювання документів виокремлюють юридичні, етичні, естетичні, політичні, релігійні, композиційні, логічні, лінгвістичні, психолінгвістичні, видавничі й поліграфічні?

А) переклад

Б) готування оглядів

В) коректура

Г) редагування.

6. Ведення діалогу – це вид

А) атрибуції текстів

Б) генерування текстів

В) розуміння текстів

Г) генерування монологічних текстів.

7. Природний інтелект – це

А) показник рівня інтелекту, розумової здатності людини, обчислюваний на основі спеціальних інтелектуальних тестів;

Б) здатність людини до розумової діяльності, або здатність до логічного мислення та прийняття певних рішень;

В) певна комп’ютерна модель природного інтелекту, або інтелекту людини;

Г) алгоритм діалогу з автоматом;

8. Термін «штучний інтелект» з’явився в науковому обігу

А) 1956 року

Б) 1955 року

В) 1954 року

Г) 1950 року.

9. У скільки етапів зводиться створення еталонного реферату?

А) 2

Б) 3
В) 4
Г) 5
10. Головні функції анотації:
А) сигнальна, пошукова
Б) довідкова, рекомендаційна
В) лінгвістична, структурна
Г) повідомлююча, перехресна.

ІІ рівень

1. Який науковець стверджує, що машина стане розумною тоді, коли буде спроможна підтримувати діалог зі звичайною людиною, а та не зможе зрозуміти, що розмовляє з машиною (діалог ведеться переписуванням)?
2. Який коефіцієнт обчислюється за формулою: r = m/Z?

ІІІ рівень

1. Що таке предметизація? У чому її завдання?
2. Відтворіть алгоритм здійснення коректури.
3. У чому основний недолік сучасних систем автоматичного реферування?

ІІІ варіант

І рівень

1. Класифікаційні індекси, дескриптори це:
А) координатне індексування
Б) інформаційно-пошукова мова
В) тезаурус
Г) авторитетні файли
2. Який вид аналітико-синтетичного опрацювання документів передбачає творчу оптимізацію:
А) реферування
Б) переклад
В) редагування
Г) коректура
3. Рівень розуміння тексту не передбачає:
А) лінгвістичний
Б) історичний
В) евристичний
Г) відтворювальний
4. Універсальна десяткова класифікація (УДК) – це приклад

такого виду аналітико-синтетичного опрацювання текстів документів, як

А) інформаційний пошук

Б) переклад

В) індексування

Г) реферування.

5. Денотативна концепція перекладу базується на тому, що

А) можна використовувати синоніми, що позначають найбільш схожі чи близькі об’єкти;

Б) перекладач проникає в семантику тексту оригіналу й виконує переклад вже саме на цій основі;

В) коли якусь інформацію не можна передати на тому лінгвістичному рівні, на якому вона є в оригіналі, її можна подати на вищому чи нижчому рівні іншими засобами цих лінгвістичних рівнів;

Г) для перекладу тексту оригіналу в мові перекладу потрібно віднайти слово, яке позначає такий самий об’єкт, як і в мові оригіналу, й замінити слово мови оригіналу словом мови перекладу.

6. Методи атрибуції досліджують і використовують у

А) палеографії та криміналістиці

Б) соціології та філософії;

В) психолінгвістиці та історіографії

Г) прикладній морфології та лексикографії.

7. Машина Тьюрінга була створена

А) 1936 року

Б) 1980 року

В) 2014 року

Г) 1913 року.

8. Тезаурусний та фреймовий підходи у системі автоматичного реферування виокремили:

А) Сєргеєв, Бєляєва

Б) Берзон, Добрускіна;

В) Волошин, Григор’єв

Г) Мазуріна, Степанчук.

9. Метод аспектного реферування \ анотування передбачає:

А) вибір речень, що мають ключові маркери;

Б) уведення частини тексту у визначений контекст;

В) творче осмислення джерела;

Г) часткову обробку тексту.

10. Вибір принципу реферування залежить від

А) стану розробленості АОТ

Б) типу тексту

В) потреб користувача

Г) всі варіанти правильні

II Рівень

1. Комплекс, що складається з бази знань і засобів виведення з них і призначений для виконання розумних рекомендацій.

2. Який параметр обчислюється за формулою: E = a/V

III рівень

1. Розкрийте практичне значення індексування.
2. Розмежуйте терміни «атрибуція» та «авторизація».
3. Яка головна мета реферування?

Модульна контрольна робота № 3

I варіант

I рівень

1. Скільки способів опису синтаксичної структури прийнято в роботах з АСА?

А) 2

Б) 3

В) 4

Г) 5

2. Редагування, за якого більшу частину операцій контролю й виправлення здійснює система редагування, а меншу – людина:

А) комп'ютеризоване

Б) автоматизоване

В) автоматичне

Г) ручне

3. Як називають виведення вихідної форми для будь-якої текстової словоформи?

А) лематизація;

Б) морфологічний аналіз

В) сегментація на морфи

Г) синтез парадигми.

4. Як називають мовний знак, за допомогою якого регулярно виражається граматичне значення?

А) машинне слово
Б) морфологічне слово
В) граматична форма
Г) суплетивна форма

5. Як по-іншому називають інтерпретаційну граматику?

А) породжувальна граматика
Б) породжувальний синтаксис;
В) теорія глибинних відмінків
Г) теорія синтаксичних категорій

II рівень

1. Синтаксичний зв’язок, при якому аналізоване слово є керувальним, головним.
2. Класичний приклад гібридної редагувально-поліграфічної системи.
3. Аналіз граматичної будови тексту, під час якого для кожної одиниці тексту визначають її граматичний клас.

III рівень

1. Як у системі АСА здійснюється контроль за дотриманням психолінгвістичних норм?
2. Проаналізуйте основні напрямки формалізації семантики.

II варіант

I рівень

1. Що є об’єктом дослідження в АСА?

А) слово
Б) словосполучення
В) речення
Г) текст

2. Редагування, за якого операції контролю і виправлення здійснює людина, комп’ютер використовують як «електронне перо»:

А) комп’ютеризоване
Б) автоматизоване
В) автоматичне
Г) ручне

3. Які види автоматичного морфологічного аналізу розрізняють?

А) зі словником основ та словником словоформ;
Б) зі словником основ, зі словником словоформ та методом

логічного множення;

В) зі словником основ, зі словником словоформ, методом логічного множення та за допомогою таблиць;

Г) за допомогою таблиць та методом логічного множення.

4. Як називають синтаксичну одиницю, меншу ніж речення, яка вільно утворюється у процесі розгортання змісту тексту та складається з одиниць мови?

А) граф

Б) семантична сітка;

В) фрейм

Г) зона зв'язку.

5. Які типи семантичних зв'язків існують у тексті?

А) морфологічні та граматичні

Б) парадигматичні та синтаксичні

В) парадигматичні та граматичні

Г) бінарні та одновалентні.

II рівень

1. Синтаксичний зв'язок, при якому аналізоване слово є керованим, залежним.

2. Метод, що задає автоматичне виправлення типових помилок за допомогою реконструюючого словника підстановок.

3. Метод встановлення синтаксичної структури речення за контактними словами, що пов'язані певним синтаксичним зв'язком.

III рівень

1. Як у системі АСА здійснюється контроль за перевіркою стилю ?

2. Яка роль представників дескриптивної школи структурної лінгвістики в розробці ідей АСА?

III варіант

I рівень

1. Який аспект вивчення речення лежить в основі АСА?

А) семантико-синтаксичний

Б) функціональний

В) комунікативний

Г) формально-синтаксичний

2. Скільки ступенів автоматизації виокремлюють у комп'ютерному редагуванні?

А) 2

Б) 3

В) 4

Г) 5

3. Редагування, за якого більшу частину операцій контролю здійснює система редагування, а меншу – людина:

А) комп’ютеризоване

Б) автоматизоване

В) автоматичне

Г) ручне

4. Яку структуру має дескрипторний словник?

А) структуру таблиці;

Б) структуру алгоритму;

В) ієрархічну структуру;

Г) структуру семантичної сітки.

5. У межах чого відбувається локальний семантичний аналіз?

А) у межах слова

Б) у межах речення;

В) у межах словосполучення

Г) у межах абзацу.

ІІ рівень

1. Програми, які дають змогу набирати, виправляти й зберігати текст.

2. Сукупність алгоритмів аналізу тексту на морфологічному, синтаксичному, логіко-семантичному рівнях його будови.

3. Метод встановлення пов’язаних синтаксичним зв’язком слів незалежно від їхнього контактного чи дистантного розташування в тексті.

ІІІ рівень

1. Як у системі АСА здійснюється контроль за перевіркою пунктуації?

2. У чому основна різниця між породжувальною та інтерпретаційною семантикою?

Модульна контрольна робота з курсу № 4

І варіант

І рівень

1. Що таке гіпертекст?

А) спосіб нелінійної подачі текстового матеріалу;

Б) ключові слова, що мають прив’язку до певних текстових фрагментів;

В) спосіб лінійної подачі текстового матеріалу;

Г) усі відповіді правильні .

2. Програма, яка дозволяє під час складання тестових завдань задавати декілька варіантів правильних відповідей?

А) Test W;

Б) MyTest;

В) Test W 2;

Г) усі відповіді правильні.

3. Найбільш важливу інформацію у слайді виділяють шрифтом розміру:

А) 18;

Б) 20;

В) 24;

Г) 30.

4. За скільки часу може бути сприйнятий ілюстративний слайд?

А) 15 секунд;

Б) 25 секунд;

В) 45 секунд;

Г) одну хвилину.

5. Словосполучення, що не можна перекласти дослівно формують:

А) автоматичні словники зворотів;

Б) ідіоматичні словники;

В) автоматичні словники фразеологізмів;

Г) автоматичні лексикологічні словники.

6. Коли активно почав розвиватися машинний переклад як специфічний напрям автоматичного опрацювання текстів?

А) в 50-х роках ХХ ст.;

Б) в 60-х роках ХХ ст.;

В) в 70-х роках ХХ ст.;

Г) в 80-х роках ХХ ст.

7. Розробником таких систем машинного перекладу, як АМПАР, СПРИНТ, НЕРПА є

А) О.Зубов;

Б) Ю.Марчук;

В) Р.Піотровський;

Г) Н.Лєонтьєва.

8. Основною одиницею перекладу, на думку Ю. Марчука, є

А) слово;

Б) речення;

В) словосполучення;

г) фраза.

9. Різновид машинного перекладу тексту, заснований на порівнянні великих обсягів мовних пар?

А) автоматичний

Б)автоматизований

В)класичний

Г)статистичний

10. Який із цих методів заснований на основі статистики?

А) інтерлінгва

Б)трансфер

В)EBMT

Г)SBMT

ІІ рівень

1. Спеціальна формально-логічна мова, яка в експліцитному вигляді подає зміст мовних одиниць та відношення між ними у вхідному тексті.

2. Концепція перекладу, що передбачала необхідність проникнення перекладача в семантику тексту оригіналу й виконання перекладу вже саме на цій основі.

3. Видавнича програма, система верстки.

ІІ рівень

1. Схарактеризуйте процес створення тестів на прикладі контрольно-діагностуючої системи Test W.

2. Письмово перерахувати та проаналізувати помилки системи МП PROMT:

Оригінал:

Nokia 9000i Communicator now supports short messages with up to 2 280 characters, the current standard being 160 characters. With the Text Web service based on Smart Messaging, the end-user is able to obtain information in a simple text format without graphics or logos from the Internet by using the short message service. Text Web information can include flight schedules, weather or traffic reports, or the stock news.

Переклад:

Nokia 9000i комунікатор тепер підтримує короткі повідомлення з до 2280 характерів (знаків), поточного стандарту, що є 160 характерами (знаками). З обслуговуванням (службою) Тканини (мережі) Тексту, заснованим на Шикарному (сильному) Messaging, кінцевий користувач здатний отримати інформацію в простому форматі тексту без графіки або емблем від Internet, використовуючи коротке обслуговування(службу) повідомлення. Інформація Тканини (мережі) Тексту може включати списки (графіки) рейса (польоту), погоду або повідомлення руху, або новини запасу (акції).

II варіант
I рівень

1. Виправлення помилок у тексті й ілюстраціях презентації здійснюється на етапі:

А) тестування;

Б) удосконалення;

В) використання;

Г) створення.

2. Скільки слайдів відводиться на навчальну інформацію:

А) 5-10;

Б) 10-15;

В) 15-20;

Г) не >5.

3. Оптимальна кількість кольорів в одному слайді:

А) 2;

Б) 3;

В) 4;

Г) більше 4.

4. Можливість безпосередньо впливати на час презентації свідчать про її:

А) багатофункціональність;

Б) мобільність;

В) ємність;

Г) інтерактивність.

5. Коефіцієнт кореляції є показником:

А) надійності тесту;

Б) валідності тесту;

В) модельності тесту;

Г) стандартності тесту

6. Хто з науковців дотримувався такої думки: «Текст перекладу має містити слова-переклади іншою мовою, оформлені синтаксично правильно, і характеризуватися стилістичною довершеністю зі збереженням авторського задуму» :

А) Р.Піотровський;

Б) О.Зубов;

В) Н.Хомський;

Г) Ю.Марчук.

7. Яка найвідоміша система машинного перекладу функціонує, починаючи з 1970-х років у США?

А) АМПАР;

Б) СИСТРАН;

В) ФРАП;

Г)ЯРАП.

8. Словником у галузі машинного перекладу називають:

А) найскладніший вид відповідностей, для якого потрібна робота спеціальних алгоритмів аналізу, синтезу і міжмовних перетворень;

Б) парадигму граматичних ознак, необхідних для встановлення відповідностей;

В) впорядкований список лексичних одиниць з необхідною інформацією;

Г) список схем і таблиць.

9.Що називають класичним методом комп`ютерного перекладу, який відбувається трьома кроками: аналіз, трансфер, генерування?

А) інтерлінгва

Б) трансфер

В) ЕВМТ

Г) SBMT

10. У якій системі ядром є блок пам`яті перекладу, в якому зберігаються речення або фрази, які часто повторюються та їх переклад?

А) інтерлінгва

Б) трансфер

В) ЕВМТ

Г) SBMT

ІІ рівень

1. Спеціальний модуль міжмовних перетворень, перезапису вхідної інформації, переносу кодів структури та лексичного наповнення вхідного тексту в коди тексту вихідної мови

2. Концепція перекладу, що базується на тому, що для перекладу тексту оригіналу в мові перекладу потрібно віднайти слово, яке позначає такий самий об'єкт як і в мові оригіналу, й замінити слово мови оригіналу словом мовою перекладу.

3. Процес додавання або віднімання простору між конкретними парами символів.

III рівень

1. Схарактеризуйте процес створення тестів на прикладі контрольно-діагностуючої системи Test W2.

2. Письмово перерахувати та проаналізувати помилки системи МП PROMT:

Оригінал:

Nokia 9000i Communicator now supports short messages with up to 2 280 characters, the current standard being 160 characters. With the Text Web service based on Smart Messaging, the end-user is able to obtain information in a simple text format without graphics or logos from the Internet by using the short message service. Text Web information can include flight schedules, weather or traffic reports, or the stock news.

Переклад з підключеним словником "Телекомунікації і зв'язок":

Nokia 9000i Комунікатор тепер підтримує короткі повідомлення з до 2280 символів, поточного стандарту, що є 160 символами. З Текстовим обслуговуванням Мережі, заснованим на Smart Messaging, кінцевий користувач здатен отримати інформацію в простому текстовому форматі без графіки або емблем від Internet, використовуючи систему передачі коротких повідомлень. Текстова інформація Мережі може включати список рейса (польоту), погоду або повідомлення трафіку, або новини фондового ринку.

III варіант

I рівень

1. Скільки етапів підготовки мультимедійної навчальної презентації виокремлюють:

А) 6;

Б) 7 ;

В) 8;

Г) 9

2. Можливість в одній презентації розмістити великий обсяг інформації свідчить про її:

А) компактність;

Б) мобільність ;

В) багатофункціональність;

Г) ємність

3. Налаштування анімаційних ефектів презентації здійснюється на етапі:

А) створення;

Б) налаштування;

В) виконання;

Г) удосконалення

4. Яка оптимальна кількість рисунків на одному слайді:

А) не >1;

Б) 1-2;

В) 2-3;

Г) 4

5. Які кольори рекомендують застосовувати у презентаціях:

А) чорно-білі;

Б) синьо-жовті;

В) зелено-сині;

Г) червоно-чорні.

6. При машинному перекладі проблема полягає у формалізації:

А) конотату і денотату;

Б) конотату і смислу;

В) повідомлення і смислу;

Г) референту і смислу.

7. Яка визначна подія, пов’язана з історією розвитку машинного перекладу, відбулась у1996 році?

А) продовжились роботи в РАН під керівництвом Ю.Апресяна;

Б) в Інституті сходознавства був розроблений ЯРАП;

В) створено АМРАП;

Г) з’явився ПЛАЙ.

8. Розробником системи СИСТРАН став:

А) П. Тома;

Б) Ю.Марчук;

В) Н.Лєонтьєва;

Г) Р.Піотровський;

9. Який метод є найпродуктивнішим для перекладу комплексних висловів?

А) інтерлінгва

Б) трансфер

В) EBMT

Г) SBMT

10. Якого з видів систем машинного перекладу не існує?

А) системи на основі граматичних правил;

Б) статистичні

В) «гібридні»

Г) семантичні.

II рівень

1.Концепція перекладу, що передбачає той факт, що коли якусь інформацію не можна передати на лінгвістичному рівні, на якому вона є в мові оригіналу, її можна подати на вищому чи нижчому рівні іншими засобами цих лінгвістичних рівнів.

2.Переклад, що передбачає скорочення оригіналу і витяг з нього найважливішої інформації і створення реферату, анотації на іншій мові.

3.Фрагмент семантичної мережі, призначений для опису деякого об’єкта предметної області з усією сукупністю властивостей; підграфи семантичних мереж; блоки знань.

III рівень

1. Схарактеризуйте процес створення тестів на прикладі контрольно-діагностуючої системи My Test.

2. Проаналізуйте та перерахуйте недоліки машинного перекладу наведеного уривку з використанням системи МП PROMT.

Переклад з використанням системи МП PROMT:

Electronic commerce over the Internet is growing at an almost exponential rate. An April 1998 report of the United States Department of Commerce, entitled The Emerging Digital Economy describes the almost mind-boggling growth of electronic commerce, and of the Internet itself. Some of the more fantastic facts included in the report are:

By the end of 1997 more than 100 million people were using the Internet and some experts expect that 1 billion people will be connected to the Internet by 2005. Traffic on the Internet is doubling every 100 days.

By 2002, Internet commerce between businesses will likely surpass $ 300 billion. The number of names registered in the domain name system grew from 26,000 in July of 1993 to 1.3 million in July of 1997.

Переклад на російську мову:

Электронная торговля по Интернету возрастает в почти показательную функцию (норма) (разряд). Апрель 1998 сообщение Отдела Соединенных Штатов Торговли, имея право Появляющейся Цифровой Экономике (экономия), описывает почти ошеломляющий рост электронной торговли и Интернета непосредственно. Некоторые из более фантастических фактов, включенных в сообщение:

К концу 1997 больше чем 100 миллионов людей использовали Интернет и некоторые эксперты ожидают, что 1 миллиард людей будет связан с Интернетом 2005. Движение в Интернете удваивается каждые 100 дней. 2002, торговля Интернета между бизнесами вероятно превзойдет 300 миллиардов. Номер (число) названий (имена), зарегистрированных в системе названия (имя) области рос от 26000 в июле от 1993 до 1.3 миллиона в июле 1997.

ЗАВДАННЯ ДЛЯ ПІДСУМКОВОЇ КОНТОЛЬНОЇ РОБОТИ

Базовий рівень

1. Наука, що вивчає системи керування:
   А) Прикладна лінгвістика
   Б) Комп’ютерна лінгвістика
   В) Кібернетика
   Г) Корпусна лінгвістика
2. Хто автор класифікації типів інформації?
   А) Партико З.
   Б) Дарчук Н.
   В) Білецький А.
   Г) Волошин В.
3. У скількох способах свого існування виявляється мова?
   А) 2
   Б) 3
   В) 4
   Г) 5
4. Наука про типи інформації:
   А) Інформатика
   Б) Кібернетика
   В) Корпусна лінгвістика
   Г) Комп’ютерна лінгвістика
5. Скільки операцій виконували комп’ютери першого покоління?
   А) 1000
   Б) 5000
   В) 10000
   Г) 100000
6. Яка елементарна база була у комп’ютерів третього покоління?
   А) Лампи
   Б) Інтегральні схеми
   В) Транзистори
   Г) Мережа транзисторів
7. Коли укладено шрифти UNICODE?
   А) 50-ті роки ХХ ст.
   Б) 70-ті роки ХХ ст.
   В) 90-ті роки ХХ ст.
   Г) п.ХХІ ст.
8. Яка одиниця виміру тривалості звука?
   А) Децибели

Б) Мілісекунди
В) Герци
Г) Хвилини

9. Для визначення тривалості звука використовують:
А) Осцилографію
Б) Спектрографію
В) Палатографію
Г) Рентгенографію

10. Які частоти не сприймає слух людини?
А) Вищі за 20 кГц
Б) Нижчі за 20 кГц
В) Вищі за 10 кГц
Г) Вищі за 5 кГц

11. Не є різновидом письма:
А) Піктографічне
Б) Ідеографічне
В) Силабічне
Г) Силабо-тонічне

12. У формулі H=G/S, G – це:
А) Кількість букв
Б) Кількість звуків
В) Кількість графем
Г) Надлишковість абетки

13. Скільки точок для кодування використовує брайлівська система?
А) 4
Б) 5
В) 6
Г) 7

14. Видом аналітико-синтетичного опрацювання текстів документів, у процесі якого перевіряють відповідність копії документа оригіналові є:
А) Реферування
Б) Редагування
В) Переклад
Г) Коректура

15. Словосполучення, що не можна перекласти дослівно формують:
А) автоматичні словники зворотів
Б) ідіоматичні словники

В) автоматичні словники фразеологізмів
Г) автоматичні лексикологічні словники

16. З якою дисципліною комп'ютерна лінгвістика пов'язана системою автоматичного перероблення тексту?
А) Лінгвістика тексту
Б) Автоматична обробка природної мови
В) Кібернетика
Г) Інформатика

17. За способом оформлення інформація буває:
А) Логічна/естетична
Б) Узуальна/спеціальна
В) Вербальна/екстравербальна
Г) Кодована/некодована

18. Скільки функцій інформації виділив Білецький А.?
А) 7
Б) 6
В) 5
Г) 3

19. Що не є об'єктом дослідження КЛ?
А) Продукт мовлення
Б) Мовлення
В) Мовна система
Г) Мовна діяльність

20. Скільки операцій виконували комп'ютери третього покоління?
А) 10000
Б) 50000
В) 100000
Г) 1000000

21. Коли почала формуватися комп'ютерна лінгвістика?
А) 50-ті роки XX ст.
Б) 80-90 ті роки XX ст.
В) 60-70 ті роки XX ст.
Г) Кінець 90-х XX ст.

22. Чим ознаменувалися 40-50-ті роки XX століття?
А) Створено перші електронні словники
Б) Здійснено синтаксичний аналізу тексту
В) Виникли комп'ютери
Г) Виникла фірма Apple

23. У чому вимірюється частота звука?

А) У герцах
Б) У мілісекундах
В) У децибелах
Г) У хвилинах

24. Для вимірювання сили артикуляції застосовують:
А) Палатографію
Б) Тензопалатографію
В) Спектрографію
Г) Осцилографію

25. Що не відноситься до фізичної характеристики звука?
А) Висота
Б) Частота
В) Інтенсивність
Г) Тривалість

26. Ієрогліфічне письмо це:
А) Піктографічне
Б) Літературно-звукове
В) Ідеографічне
Г) Складове

27. Який компонент не становить систему письма?
А) Напрям письма
Б) Орфоепія
В) Орфографія
Г) Графіка

28. У формулі A=L/S, S – це:
А) Кількість букв
Б) Кількість графем
В) Кількість звуків
Г) Оптимальність абетки

29. Який спосіб кодування традиційного письма застосовують в електронних каналах звука?
А) Шифрування
Б) Азбука Морзе
В) Криптографування
Г) Шрифт Брайля

30. Встановлення справжнього автора тексту документу це:
А) Огляд
Б) Коректура
В) Генерування

Г) Атрибуція

Середній рівень

1. Назвіть вид спеціального письма, за допомогою якого фіксують вимову усних текстів.
2. Назвіть вид спеціального письма, за допомогою якого тексти, написані однією графікою, політерно відтворюють за допомогою іншої графіки.
3. Назвіть вид спеціального письма, за допомогою якого усний текст можна записувати в кілька разів швидше, ніж за допомогою звичайного письма.
4. Назвіть вид спеціального письма, який дає змогу записати зафіксований природною писемною мовою текст в такий спосіб, який максимально ускладнює його політерне розпізнавання.
5. Назвіть процес виправлення граматичних і технічних помилок та недоліків, як в текстовому, так і графічному матеріалах, підготовлених для розмножения друкарським (або будь-яким іншим) способом.
6. Назвіть процес визначення достовірності, автентичності художнього твору, його автора, місця й часу створення на підставі аналізу стилістичних і технологічних особливостей.

Високий рівень

1. Оберіть види аналітико-синтетичної обробки інформації:
   А) бібліографічний опис
   Б) генерування
   В) індексування
   Г) атрибуція
   Д) переклад
2. Оберіть види індексування:
   А) систематизація
   Б) предметизація
   В) координатне індексування
   Г) вільне індексування
   Д) контрольоване індексування
3. Які ознаки має реферат:
   А) наявність чіткої авторської позиції
   Б) максимально стислий обсяг
   В) зміст реферату повністю залежить від змісту реферованих джерел
   Г) містить точний виклад основної інформації без спотворень і

суб’єктивних оцінок
Д) має постійні структури
4. Оберіть види редагування:
А) Наукове
Б) Загальне
В) Галузеве
Г) Спеціальне
Д) Творче
5. За рівнем узагальнення інформації та впливом на аудиторію розрізняють огляди:
А) Інформаційні
Б) Політичні
В) Аналітичні
Г) Універсальні
Д) Тематичні
6. Набір смислових частин діалогу вміщує:
А) Підготовка до розмови
Б) Установлення контакту
В) Початок розмови
Г) Розвиток теми
Д) Завершення розмови
7. Інформація може бути представлена у формах:
А) Знаково-письмовій
Б) Буквеній
В) Жестів або сигналів
Г) Словесній формі
Д) Паперовій
8. За носієм та засобами відтворення автоматичні або електронні словники поділяються на:
А) Паперові
Б) Комп’ютерні
В) Кишенькові
Г) Мобільні
Д) Аудіальні
9. Оберіть вимоги, що вливають на ефективність лінгвістичних баз даних:
А) значна кількість даних
Б) незначна кількість даних
В) відкритий доступ до даних

Г) анонімність
Д) гарантована відсутність збоїв.

10. Оберіть складники інтелектуальної діяльності людини:
А) Пізнання
Б) Пам’ять
В) Увага
Г) Розуміння
Д) Генерування знань

11. Оберіть підходи до створення систем штучного інтелекту:
А) Історичний
Б) Логічний
В) Структурний
Г) Еволюційний
Д) Імітаційний

12. З якими нелінгвістичними дисциплінами пов’язана комп’ютерна лінгвістика?
А) Географія
Б) Політологія
В) Археологія
Г) Біологія
Д) Інформатика

13. Оберіть основні типи лінгвістичних моделей:
А) Описова / історична
Б) Індуктивна / дедуктивна
В) Мовна / мовленнєва
Г) Оригінальна / перекладна
Д) Статична / динамічна

14. Оберіть способи вираження граматичного значення:
А) Семантичний
Б) Синтетичний
В) Аналітичний
Г) Аналітико-синтетичний
Д) Логічний

15. Вертикальна класифікація комп’ютерної лінгвістики вміщує рівні:
А) Прагматика
Б) Дериватологія
В) Стилістика
Г) Семантика

Д) Граматика

16. Об’єктами дослідження комп’ютерної лінгвістики є:

А) Мовний продукт

Б) Мова загалом

В) Мовна система

Г) Мовлення

Д) Мовна діяльність

17. Якими факторами зумовлений вибір принципів комп’ютерного морфологічного аналізу:

А) Авторством тексту

Б) Тематикою тексту

В) Системою мови

Г) Системою письма і друку

Д) Закономірностями породження мовлення

18. Оберіть типи автоматичного морфологічного аналізу залежно від характеру його лінгвістичного забезпечення та способу розпізнавання морфологічного структури тексту:

А) Автоматичний морфологічний аналіз зі словником основ слів

Б) Автоматичний морфологічний аналіз зі словником словоформ

В) Автоматичний морфологічний аналіз методом логічного множення

Г) Автоматичний морфологічний аналіз із перекладними та тлумачними словниками

Д) Автоматичний морфологічний аналіз за допомогою таблиць

19. Оберіть типи автоматичного синтаксичного аналізу залежно від сфери його застосування, вихідних елементів та способів виконання:

А) Універсальні та часткові системи, придатні для розв’язання окремих дослідницьких завдань

Б) Спеціальні системи, що генерують складне синтаксичне ціле

В) Системи, що моделюють тексти

Г) Системи, що встановлюють синтаксичні структури в тексті за частинами мови словоформ та за їхніми синтаксичними ролями

Д) Системи з безперервним та циклічним переглядом тексту

20. Оберіть стратегії (методики) для розробленні систем автоматичного синтаксичного аналізу:

А) Послідовний аналіз тексту

Б) Передбачувальний аналіз

В) Методика опорних точок

Г) Методика дослідження синтаксичної ролі слів у реченнях
Д) Методика фільтрів

21. Оберіть ознаки, на основі яких здійснено класифікації систем машинного перекладу в комп’ютерній лінгвістиці:

А) Підхід до опрацювання тексту
Б) Специфіка вхідної та вихідної мов
В) Стилістика тексту
Г) Спосіб зіставлення текстів вхідною та вихідною мовами
Д) Роль людини у здійсненні машинного перекладу

ПИТАННЯ ДО ІСПИТУ

1. Об’єкт, предмет, мета та завдання курсу квантитативна та комп’ютерна лінгвістика.
2. Місце квантитативної та комп’ютерної лінгвістики у мовознавстві.
3. Місце квантитативної та комп’ютерної лінгвістики у кібернетиці. Взаємозв’язок комп’ютерної лінгвістики з іншими науками.
4. Етап виникнення комп’ютерних лінгвістичних систем.
5. Етап експериментальних комп’ютерних лінгвістичних систем.
6. Етап промислових комп’ютерних лінгвістичних систем колективного користування.
7. Етап промислових комп’ютерних лінгвістичних систем індивідуального користування.
8. Етап комп’ютерних лінгвістичних мереж.
9. Сучасний стан комп’ютерної лінгвістики в Україні.
10. Прикладна фонетика: мовний звук; частота звука; інтенсивність; тривалість; спектральний склад звуків; інструменти аналізу звуків.
11. Традиційні системи письма: види систем письма; критерії оцінювання графіки; пристрої для традиційного письма.
12. Транскрибування: звукове транскрибування, інтонаційне транскрибування. Застосування транскрибування.
13. Транслітерування.
14. Стенографування.
15. Системи письма для незрячих. Мова жестів.
16. Криптографування. Азбука Морзе. Цифрові системи письма.
17. Поняття про аналітико-синтетичне опрацювання документів.
18. Індексування.
19. Інформаційний пошук.
20. Коректура. Редагування.
21. Готування оглядів.
22. Атрибуція текстів.
23. Генерування текстів. Ведення діалогу.
24. Складові штучного інтелекту.
25. Підходи до створення систем штучного інтелекту.
26. Машина та тест Тюрінга.
27. Визначення можливості створення штучного інтелекту.

28. Місце лінгвістичного забезпечення в системах штучного інтелекту загального призначення.
29. Складність створення систем штучного інтелекту.
30. Передумови досліджень в галузі автоматичного синтаксичного аналізу (АСА).
31. Особливості здійснення процедури АСА.
32. Роль дистрибутивного методу і методу безпосередніх складників (БС) у розробці (АСА).
33. Граматика залежностей як спосіб представлення синтаксичної структури речення.
34. Дерева залежностей у системах АСА.
35. Принципи роботи деяких синтаксичних аналізаторів: системи АОТ, Dictum, аналізатор на основі системи SMART.
36. Текстові процесори.
37. Системи редагування. Технологічні особливості комп'ютерного редагування.
38. Категоризація лексики та семантичні характеристики слів, що застосовуються при автоматичному семантичному аналізу мови.
39. Модель «м'якого» розпізнавання тексту комп'ютером.
40. Методики визначення в тексті ключових слів (слів-концептів).
41. Автоматичне індексування тексту.
42. Різновиди систем інформаційного пошуку (ІПС).
43. Інформаційно-пошукові мови (ІПМ): класифікаторні та дескрипторні.
44. Інформаційно-пошукові тезауруси (ІПТ).
45. Машинне реферування й анотування тексту шляхом його стиснення.
46. Тезаурусне реферування науково-технічного тексту.
47. Об'єктивні оцінки якості реферування. Процедура створення еталонного реферату.
48. Реферування на основі пофразної семантичної мережі.
49. Основні види реферування, анотування тексту.
50. Підходи до створення реферату, анотації.
51. Основні методи, на яких базується лінгвістичний аналіз побудови реферату, анотації.
52. Критика штучного розуму.
53. Поняття експертних систем, їх характеристика.
54. Лінгвістичні засади створення автоматичного морфологічного аналізу (АМА).

55. Експериментальні та промислові системи АМА.
56. Типологія алгоритмів автоматичного морфологічного аналізу.
57. Принципи роботи деяких аналізаторів (АОТ, Mocky, Mystem).
58. Історія машинного перекладу.
59. Принципи машинного перекладу. Лінгвістичне забезпечення систем машинного перекладу.
60. Структура системи машинного перекладу. Алгоритми машинного перекладу.
61. Системи машинного перекладу та їх класифікація.
62. Системи прямого перекладу.
63. Системи перекладу за принципом трансферу.
64. Системи перекладу з мовою-посередником.
65. Модель машинного перекладу на основі перекладних відповідників АМПАР.
66. Система машинного перекладу СИСТРАН.
67. Джорджтаунська система машинного перекладу.
68. Система РУТА.
69. Типові помилки машинного перекладу та оцінка якості перекладу.
70. Проблеми комп'ютерної лексикографії.
71. Вимоги до систем машинного перекладу, нерозв'язані проблеми машинного перекладу.
72. Основи і правила комп'ютерного набору тексту.
73. Характеристика комп'ютерних програм Adobe PageMaker, Adobe InDesign.
74. Створення тестів на прикладі програми Test W, Test 2W, My Test.
75. Мультимедійна презентація як сучасний засіб упровадження інформаційних технологій. Вимоги до презентацій.

СЛОВНИК ТЕРМІНІВ КУРСУ

Автоматична обробка природної мови – напрям комп'ютерної лінгвістики, який передбачає створення, перетворення й аналіз текстів із застосуванням природної або штучної мов, результатом чого може бути формування машинних фондів національних мов, автоматичних словників, термінологічних банків, комп'ютерних картотек, баз даних, комп'ютерних граматик тощо. **Automated language processing, language data processing**

Автоматичне розпізнавання звукового мовлення – напрям прикладної лінгвістики, зокрема, її фонетичної галузі, метою якого є створення систем штучного інтелекту, здатних розпізнавати звукові повідомлення. **Automatic speech recognition**

Автоматичний морфологічний аналіз тексту – аналіз граматичної будови тексту, під час якого для кожної одиниці тексту визначають її граматичний клас та підкласи.

А**втоматичний семантичний аналіз** – сукупність методів і прийомів, за допомогою яких можна шляхом однозначних формальних процедур за правилами певної формалізованої граматики, що реалізується на комп'ютері за спеціальними лінгвістичними алгоритмами, з досить високою точністю однозначно представити смисл.

Автоматичний синтаксичний аналіз тексту – аналіз граматичної будови тексту, спрямований на встановлення типів синтаксичних структур речень, визначення синтаксичних ролей в них окремих слів та з'ясування типів синтаксичних зв'язків як між словами всередині речень, так і між реченнями.

Автоматичний синтаксичний аналіз за безпосередніми складниками речення – метод встановлення синтаксичної структури речення за контактними словами, що пов'язані певним синтаксичним зв'язком.

Автоматичний синтаксичний аналіз за зв'язком залежності між словами – метод встановлення пов'язаних синтаксичним зв'язком слів незалежно від їхнього контактного чи дистантного розташування в тексті.

Автоматичний синтез звукового мовлення – напрям прикладної лінгвістики (зокрема, фонетики), метою якого є комп'ютерне програмування людьми породження усного звукового мовлення системами штучного інтелекту. **Automatic speech**

synthesis

Аналітико-синтетичне опрацювання документів – це процеси перетворення інформації, що міститься в первинному документі, з метою створення вторинних документів.

Афективна сюжетна одиниця – елемент структури оповідальних текстів, що застосовується в комп'ютерній обробці тексту для опису когнітивно значимого зв'язку між позитивно й негативно оцінюваними станами персонажів тексту. **Affective plot unit**

База даних – структурований і формалізований масив інформації про певну предметну галузь. **database**

База знань – структурований і формалізований набір відомостей про об'єкти певної предметної галузі та відношення між ними, на основі яких можна будувати судження про них, здійснювати різноманітні операції логічних умовиводів. **knowledge base**

Гіпертекст (від гр. hyper – понад і текст) – 1) особливий метод побудови інформаційних систем, що забезпечує прямий доступ до інформації на підставі логічного зв'язку між її блоками; 2) система представлення текстової та мультимедійної інформації у вигляді мережі пов'язаних між собою текстових й ін. файлів, яка застосовує нелінійний, асоціативно-фрагментарний і сітковий принципи репрезентації інформації. **Hypertext**

Граф – 1) у графології варіант графеми (алограф); 2) у когнітивній лінгвістиці спосіб графічного представлення зв'язку між значеннями слова-полісеманта (семантичний граф), між компонентами семантичної або синтаксичної структури речення, предикатно-аргументної структури, пропозиції, а також між фреймами-схемата і схемами (концептуальні графи). **Graph**

Дескриптори – ключові слова текстового документа, сукупність яких є його пошуковим позначенням, тобто використовується для швидкого й ефективного пошуку необхідної інформації у системах штучного інтелекту. **Descriptors**

Доморфологічний аналіз – етап, на якому в тексті визначають одиниці, для розпізнавання властивостей яких не потрібні процедури автоматичного морфологічного аналізу.

Інтерфейс – комплекс лінгвістичних і нелінгвістичних засобів, спрямований на підтримання діалогу користувача з комп'ютерною програмою. **Interface**

Інформаційно-пошукова мова – штучна формалізована мова, призначена для лінгвістичного забезпечення інформаційно-пошукових систем (для представлення змісту документів і запитів щодо пошуку інформації), яка містить систему символів (алфавіт, цифрові позначення, пунктуаційні знаки, спеціальні позначки), граматичні правила парадигматики й синтагматики одиниць, правила перекладу, правила використання, словники. **Information query language, data-base query language**

Інформаційно-пошукова система – комплекс пов'язаних між собою частин текстів, призначених для пошуку й вияву елементів інформації, які є відповіддю на інформаційний запит, пред'явлений системі**. Information(al) retrieval System**

Інформація (від лат. informare – зображувати, повідомляти) – сукупність знань, образів, відчуттів, наявних у свідомості людини або штучному інтелекті, які поступають по різних каналах передачі, переробляються й використовуються у процесі життєдіяльності людини й роботі автоматичних комп'ютерних систем. **Information**

Комп'ютерна морфологія – частина комп'ютерної граматики, яка моделює творення форм слова (словозміну) або окремих слів (словотворення).

Комбінаторний вибух – різке зростання кількості можливих варіантів аналізу речення та його окремих частин при комп'ютерній обробці у програмах, які застосовують традиційний рівневий підхід: від фонетики, морфології до синтаксису й семантики. **Combinatorial explosion**

Комп'ютерна граматика – сукупність алгоритмів аналізу тексту на морфологічному, синтаксичному та логіко-семантичному рівнях його будови, втілених у таблицях, графах, матрицях, зведеннях правил.

Комп'ютерна лінгвістика – маргінальна галузь мовознавства, спрямована на розробку автоматизованих методів зберігання, обробки, переробки й використання лінгвістичних знань й інформації, репрезентованої знаками природної мови. **Computational linguistics**

Комп'ютерна семасіологія – розділ комп'ютерної граматики, який моделює зміст окремих мовних одиниць та тексту в цілому.

К**омп'ютерний синтаксис –** розділ комп'ютерної граматики, який моделює будову речень та складних синтаксичних цілих.

Коректура – це вид аналітико-синтетичного опрацювання

текстів документів, у процесі якого перевіряють відповідність копії документа його оригіналові.

Корпусна лінгвістика – галузь прикладного мовознавства, яка займається формуванням комп'ютерних корпусів текстів у різних мовах і спрямована на максимально об'єктивний аналіз мовних явищ в умовах реальної живої комунікації. **Corpus linguistics**

Корпусний аналіз – один з об'єктивних методів мовного аналізу, спрямований на вивчення певних закономірностей та особливостей мови й мовлення на підставі обробки комп'ютерного корпусу текстів. **Corpus method**

Лематизація (від лат. lemma-словникова форма) – процедура комп'ютерної й корпусної лінгвістики, що відновлює словникову форму за її словоформою. **Lemmatization**

Лінгвістична інтелектуальна комп'ютерна система – організована сукупність процедур для моделювання розумової діяльності людини в процесі розв'язання теоретичних або практичних завдань опрацювання мовної інформації.

Лінгвістичний аналізатор (парсер) – комп'ютерна програма, яка аналізує граматичну структуру речення. **Linguistic parser**

Лінгвістичний процесор – багаторівнева система автоматичної обробки природної мови (аналізу й синтезу текстів), здатна забезпечувати діалог у системах штучного інтелекту. **Natural language processor**

Математична лінгвістика – маргінальний розділ мовознавства, математики, семіотики й логіки, спрямований на розроблення формального апарату, способів математичного та статистичного аналізу різних мовних явищ і дослідження загальних законів побудови знакової системи, якою є природна мова, шляхом математичного моделювання. **Mathematical linguistics**

Машинний (автоматичний) переклад – галузь прикладної лінгвістики, спрямована на комп'ютерне програмування та створення автоматичних систем перекладу з однієї мови на іншу переважно науково-технічних та ділових текстів. **Machine translation**

Метамова – універсальний формалізований семіотичний код семантичного й синтаксичного опису природних мов. **Metalanguage**

Методика семантичного інтеграла (атрибуції) – одна з

експериментальних методик психолінгвістики, спрямована на встановлення авторства тексту (атрибуції). **Method of semantic integral**

Мова програмування – система спеціальних знакових засобів, що дає змогу записувати завдання для ЕОМ із метою їхнього виконання комп'ютером і забезпечує його інформаційну взаємодію з людиною. **Programming language**

Мова-еталон – метамова опису природної мови або спосіб аналізу систем порівнюваних природних мов зі своєю системою й параметрами. **Ideal language type**

План – у штучному інтелекті й комп'ютерній лінгвістиці сукупність процедур, породжених сценарієм, які сприяють досягненню мети. **Plan**

Предикатно-аргументна структура – термін семантичного синтаксису й ситуаційного моделювання; глибинний синтаксичний аналог судження, центром якого є предикат, що корелює з аргументами й ад'юнктами. **Predicate-argument structure**

Преференційна семантика – семантична концепція, яка застосовується при автоматичній обробці природних мов у комп'ютерній лінгвістиці та ґрунтується на зіставленні кожного знака з певною семантикою за умови фіксації його переважних змістовних зв'язків з іншими знаками й мінімізації ролі синтаксису. **Preference semantics**

Прикладна фонетика – маргінальна галузь прикладної лінгвістики й фонетики, яка застосовує знання звукових механізмів мовлення, фонологічної системи мови, процедур сприйняття звукового мовлення для розв'язання різних прикладних завдань, як-от: навчання читанню, письму; вивчення іноземних мов, кодифікації норм вимови та правопису, усунення дефектів звукового мовлення, автоматичного синтезу й розпізнавання мовлення тощо. **Applied phonetics**

Процедурна семантика – напрям лінгвістичної семантики, який розробляє метамову опису значення тексту чи речення та встановлює сукупність інструкцій як операцій над знаками, що сприяють уведенню нового знання до моделі світу реципієнта. **Procedural semantics**

Редагування – це вид аналітико-синтетичного опрацювання текстів документів, у процесі якого текст документа приводять у відповідність до існуючих норм, а також здійснюють його творчу

оптимізацію.

Реферуванням документа – метод мікроаналітичного згортання інформації з первинного документа наукового або виробничо-практичного професійного характеру, а також сам процес створення реферату як вторинного документа.

Семантична сітка – формалізована модель представлення значення мовних одиниць, конкретної ситуації або їхньої сукупності у вигляді мережі зв'язків між елементами (вузлами) – значеннями чи концептами та їхніми знаками як складниками відповідних ситуацій. **Semantic net (work)**

Система автоматичного перероблення тексту – лінгвістична інтелектуальна комп'ютерна система для виконання процедур морфологічного, синтаксичного та логіко-семантичного аналізу тексту.

Система інтелектуальної підтримки – у комп'ютерній лінгвістиці клас систем штучного інтелекту, які моделюють окремі інтелектуальні здатності людини і використовують їх із метою прийняття рішень, часткового контролю, інформування й аналізу. **Intelligent assistance System**

Скрипт – структура репрезентації процедурних знань (як і сценарій), що стосуються конкретних учасників ситуації (наприклад, обслуговування у магазині). **Script**

Структури репрезентації знань – умовні та спрощені моделі свідомості, які представляють різнопланову інформацію, набуту на підставі різних пізнавальних психічних механізмів людини в її взаємодії з навколишнім світом й у процесі внутрішньої рефлексії. **Knowledge structure, mental representation**

Схема – структура репрезентації процедурних знань, альтернативна фрейму, який представляє декларативні знання. **Scheme**

Схема синтезу тексту – модель комп'ютерного породження (синтезу) описового тексту, яка передбачає чотири головних схеми його аналізу: атрибутивну – зразок приписування властивостей, якісних ознак різним об'єктам; конститутивну, яка зумовлює опис текстом об'єктів у термінах їхніх складників і різновидів; ідентифікаційну, спрямовану на встановлення класу, до якого належать об'єкти, й контрастивну, яка визначає негативну протилежність головного уявлення тексту. **Scheme oftext synthesis**

Сценарій – структура репрезентації процедурних знань, яка

представляє неавтоматичну життєву ситуацію та правила цілеспрямованої поведінки в ній, на відміну від фрейму, що оперує декларативними знаннями. **Scenario**

Теорія концептуальних залежностей – концепція американського дослідника Р. Шенка, згідно з якою базою породження й розуміння текстів є концептуальні залежності. **Conceptual Dependency Theory**

Теорія міжмовних відповідників – концепція встановлення закономірної відповідності між одиницями оригіналу й перекладу на системному й функціональному рівнях.

Тест Тьюринга – мисленнєвий експеримент англійського математика А. Тьюринга, що ґрунтувався на питанні про можливість наявності інтелекту в комп'ютера. **Turing test**

Трансфер – стратегія побудови програм машинного перекладу за допомогою введення проміжної мови-посередника. **Transfer**

Фрейм (від англ. framework – каркас) – структура репрезентації знань, у якій відображено набуту досвідним шляхом інформацію про деяку стереотипну ситуацію та про текст, що її описує, а також інструкцію по її використанню. **Frame**

БІБЛІОГРАФІЯ

Базова

1. Баранов А. Н. Введение в прикладную лингвистику / А. Н. Баранов. – М. : Эдиториал УРСС, 2001. – 345 с.

2. Волошин В. Г.Комп'ютерна лінгвістика: Навчальний посібник. / В. Г. Волошин. – Суми : ВТД «Українська книга», 2004. – 382 с.

3. Дарчук Н. П. Комп'ютерна лінгвістика (автоматичне опрацювання тексту): підручник / Н. П. Дарчук. – К. : Видавничо-поліграфічний центр «Київський університет», 2008. – 351 с.

4. Карпіловська Є. А. Вступ до комп'ютерної лінгвістики / Є. А. Карпіловська. – Донецьк : Юго-Восток, 2003. – 184 с.

5. Карпіловська Є. А. Вступ до прикладної лінгвістики: комп'ютерна лінгвістика : Підручник. / Є. А. Карпіловська – Донецьк : ТОВ «Юго-Восток, ЛТД», 2006. – 188 с.

6. Марчук Ю. Н. Компьютерная лингвистика / Ю. Н. Марчук. – АСТ : Восток-Запад, 2007. – 317 с.

7. Партико З. В. Прикладна і комп'ютерна лінгвістика : Вступ до спеціальності: Навчальний посібник./ З. В. Партико – Л. : Афіма, 2008. – 224 с.

8. Пещак М. М. Нариси з комп'ютерної лінгвістики / М. М. Пещак. – Ужгород : Закарпаття, 1999. – 199 с.

Допоміжна

1. Александрова З. В. Мультимедійна презентація як сучасний засіб навчання / З.В. Александрова [Електронний ресурс].

2. Апресян Ю. Д. О языке толкований и семантических примитивах / Ю. Д. Апресян // Логический анализ языка. Ментальные действия. – М. : Рос. АН, Ин-т языкознания, 1993. – С. 7-22.

3. Бірюков А. В. Тестування лінгвістичної якості систем машинного перекладу «Плай», «Прагма» та «Ленгвідж майстер» / А. В. Бірюкова [Електронний ресурс].

4. Блехман М. С. Машинный перевод: история и реалии / М. С. Блехман // Компьютерное обозрение. – 1996. – № 5(29). – С. 22–23.

5. Бук С. Н. Основи статичної лінгвістики: Навчально-методичний посібник/ Відп. ред. проф. Ф. С. Бацевич. –

Видавничий центр ЛНУ імені Івана Франка, 2008. – 124 с.
6. Використання інтерактивних методів та мультимедійних засобів у підготовці педагога : [зб. наук. праць] / Кам'янець-Подільський держ. ун-т / В. М. Федорчук (відп. ред.). – Кам'янець-Подільський : Абетка-НОВА, 2003. – 208 с.
7. Волошиновська І. А. Ефективність авторської та тематичної атрибуції текстів науково-технічного спрямування / І. А. Волошиновська // Лінгвістичні студії : [зб. наук. пр.]. – Донецьк : ДонНУ, 2011. – № 23. – С. 242–247.
8. Городенко Л. Системи верстки: Практ. посіб. для студ. / Л. Городенко – К. : Центр вільної преси, 2006. – 520 с.
9. Городецький Б. Ю. Комп'ютерна лінгвістика: моделювання мовного спілкування // Нове в зарубіжній лінгвістиці. – Вип. XXIV. Комп'ютерна лінгвістика. – М., 1989. – С. 5–31.
10. Грязнухіна Т. О., Нікула М. В. Система автоматичного морфологічного аналізу українського наукового тексту / Т. О. Грязнухіна, М. В. Нікула // Проблеми українізації комп'ютерів. Матеріали 2-ї міжнародної конференції. – Київ, 1993. – С. 42–46.
11. Гуревич Р. С. Інформаційні технології навчання : інтегрований підхід / Р. С. Гуревич, М. Ю. Кадемія, М. М. Козяр. – Львів : Вид-во «СПОЛОМ», 2011. – 484 с.
12. Дарчук Н. Автоматичний синтаксичний аналіз текстів корпусу української мови / Н. Дарчук // Українське мовознавство. – 2013. – Вип. 43. – С. 11–19.
13. Дарчук Н. П. Сегментация сложного предложения на составляющие предикативные части / Н. П. Дарчук // Синтаксический анализ научного текста на ЭВМ. – Киев, 1999. – С. 40–127.
14. Дарчук Н. П. ЭВМ в синтаксических исследованиях / Н. П. Дарчук // Использование ЭВМ в лингвистических исследованиях. – Киев, 1990. – С. 113–129.
15. Дарчук Н. П. Напрямки формалізації семантики / Н. П. Дарчук // Мовні і концептуальні картини світу. – 2013. – Вип. 46(1). – С. 385–396.
16. Дибкова Л. М. Інформатика і комп'ютерна техніка : Навчальний посібник. Видання 2-ге, доповнене. – К. : Академвидав, 2005. – 416 с.
17. Евдокимов А. С. Искусство машинного перевода. /

А. С. Евдокимов – М. : Hard’N’Soft, 2005. – №7. – С. 86–91.
18. Захарова И. Г. Информационные технологии в образовании : Учеб. пособие для студ. высш. пед. учеб. заведений. – М. : Издательский центр «Академия», 2003. – 192 с.
19. Іванов В. Ф. Комп’ютерні мас-медіа на межі століть / В. Ф. Іванов // Актуальні питання масової комунікації. – Вип. 3. – К. : Ін-т жур-ки, 2002. – С. 41–42
20. Карпенко В. О. Основи редакторської майстерності. Теорія, методика, практика: Підручник. / В. О. Карпенко – К. : Університет «Україна», 2007. – 431 с.
21. Карпенко О. О. Аналітико-синтетична переробка документної інформації: бібліографічний опис : навч. посіб. / О. О. Карпенко. – 2-ге вид., доповн. і переробл. – Х. : Нац. аерокосм. ун-т «Харк. авіац. ін-т», 2010. – 138 с.
22. Кісіль М. В. До проблеми машинного перекладу / М. В. Кісіль // Актуальні проблеми соціальної комунікації : збірник матеріалів конференції. – Івано-Франківськ : В-во ІФНГУНГ, 2011. – С. 97–103.
23. Костриба Л. О. Створення мультимедійних програмно-педагогічних систем навчання засобами Powerpoint / Л. О. Костриба // Комп’ютер у школі та сім’ї. – 2008. – № 4. – С. 33–36.
24. Кузьменко Д. М. Комп’ютерна лінгвістика і Web 2.0. / Д. М. Кузьменко – К. **:** Наукова думка, 2009. – 250 с.
25. Лангенбах М. Автоматичний синтаксичний аналіз речення за принципами граматики залежностей / М. Лангенбах // Науковий вісник Східноєвропейського національного університету імені Лесі Українки. Серія : Філологічні науки. Мовознавство. – Луцьк, 2015. – № 3. – С. 249–254.
26. Левицкий В. В. Квантитативные методы в лингвистике / В. В. Левицкий. – Винница : Нова Книга, 2007. – 213 с.
27. Лутовинова В. І. Реферування як процес мікроаналітичного згортання інформації : практ. посіб. / В. І. Лутовинова. – К. : АПН України, 2007. – 73 с.
28. Мацак Ж. Г., Скоробогатова Т. І. Машинний переклад та його специфіка / Ж. Г. Мацак, Т. І. Скоробогатова [Електронний ресурс].
29. Міщенко А. Л. Сучасні методи, напрямки й здобутки комп’ютерної лінгвістики / А. Л. Міщенко // Наукові записки Кіровоградського державного педагогічного університету імені В.Винниченка. – Вип. 95 (2). – Серія: Філологічні науки

(мовознавство). – Кіровоград : РВВ КДПУ ім. В.Винниченка, 2011. – С. 122–233.
30. Мокрогуз О. П. Мультимедійна презентація в системі засобів навчання / О. П. Мокрогуз // Комп'ютер у школі та сім'ї. – 2009. – № 8. – С. 21–23.
31. Партико З. В. Галузеве редагування в засобах масової інформації: Конспект лекцій. / З. В. Партико – Львів : Афіша, 2007. – 104 с.
32. Партико З. В. Загальне редагування: нормативні основи: Навч. посіб. /З. В. Партико – Львів : Афіша, 2006. – 416 с.
33. Пиотровский Р. Г. Текст, машина, человек. / Р. Г. Пиотровский – М. : Наука, 1987. – 160 с.
34. Прохорова Е. А. Компьютер как инструмент. / Е. А. Прохорова – К. : Випол, 2000. – 198 с.
35. Рыбаков Ф. И. Автоматическое индексирование на естественном языке / Ф. И. Рыбаков, Е. А. Руднев, В. А. Петухов. – М. : Энергия, 1980. – 160 с.
36. Селіванова О. О. Актуальні напрями сучасної лінгвістики (аналітичний огляд) / О. О. Селіванова. – К. : Український фітосоціологічний центр, 1999. – 148 с.
37. Селіванова О. О. Лінгвістична енциклопедія / О. О. Селіванова. – Полтава : Довкілля-К, 2010. – 844 с.
38. Хвощ П. М. Прикладна лінгвістика сьогодні / П. М. Хвощ. – К. : Світ, 2004. – 210 с.
39. Цихоцька Я. О. Автоматичний синтаксичний аналіз / Я. О. Цихоцька // Матеріали доповідей V Регіональної студентської наукової конференції «Мова та література як об'єкти філологічного дослідження» 21 березня 2012 р., м. Харків. – Х. : ХНУ ім. В. Н. Каразіна, 2012. – С. 120–121.
40. Шевченко І. А. Використання мультимедійних презентацій як сучасного засобу навчання в системі післядипломної освіти / І. А. Шевченко // Комп'ютер у школі та сім'ї – №1, 2012 – С. 23–26.
41. Шлыкова О. В. Культура мультимедиа : учеб. пособие [для студ. вузов] / О. В. Шлыкова / Московский гос. ун-т культуры и искусств. – М. : Фаир-Пресс, 2004. – 415 с.