

УДК 519.62



ПОБУДОВА ФУНКЦІОНАЛЬНИХ МОДЕЛЕЙ ЕЛЕМЕНТІВ КОМП'ЮТЕРНОЇ ЛІНГВІСТИКИ

І.Д. Вечірська¹, Т.М. Федорова², Г.Г. Четвериков³

¹ ХНУРЕ м. Харків, Україна, ira_se@list.ru

² ХНУРЕ м. Харків, Україна, tanja_fyodorova@mail.ru

³ ХНУРЕ м. Харків, Україна, chtvergg@kture.kharkov.ua

Стаття присвячена аналізу структури комп'ютерної лінгвістики та побудові функціональних моделей для опису задач та процесів обробки мовної інформації в корпусній лінгвістиці, машинному перекладі та системах пошуку та класифікації. Наведені приклади автоматичного розпізнавання символів, автоматичного реферування текстів, створення електронних словників, тезаурусів, онтологій, автоматичного розпізнавання мови і синтезу мовлення, оптичного розпізнавання символів.

ЛІНГВІСТИКА, КОМП'ЮТЕРНА ЛІНГВІСТИКА, КОРПУС, МАШИННИЙ ПЕРЕКЛАД, ІН-
ФОРМАЦІЙНО-ПОШУКОВА СИСТЕМА, СЛОВНИК, МОДУЛЬ, АНАЛІЗ

Вступ

Мова є найважливішим засобом комунікації в суспільстві і тісно пов'язана з мисленням і свідомістю. Мовознавство являється однією з центральних наук і входить до кола гуманітарних наукових дисциплін, які досліджують людину і людське суспільство. Лінгвістика вивчає не тільки існуючі (існуючі або можливі в майбутньому) мови, а й людську мову взагалі. Лінгвістику можна розділити на наукову (яка передбачає побудову лінгвістичних теорій) і практичну. Найчастіше під лінгвістикою мають на увазі саме наукову лінгвістику [1].

Лінгвістика безпосередньо вивчає лише факти мови або мовні явища, тобто мовні акти носіїв живої мови разом з їх результатами (текстами) або мовний матеріал (обмежена кількість письмових текстів на мертвій мові, якою вже ніхто не користується як засобом спілкування).

Лінгвістика займається не тільки спостереженням, реєстрацією та описом фактів мови, а й висуненням гіпотез для пояснення цих фактів, формулюванням гіпотез у вигляді теорій і моделей, що описують мову, їх експериментальною перевіркою і спростуванням; прогнозуванням мовної поведінки. Пояснення фактів буває внутрішнім (через мовні ж факти) або зовнішнім (через факти фізіологічні, психологічні, логічні або соціальні).

Лінгвістику в широкому сенсі слова (пізнання мови і передача результатів цього пізнання іншим людям) можна розділити на наступні напрямки:

- теоретична лінгвістика (наукова) передбачає побудову лінгвістичних теорій, розглядає всі аспекти і проблеми, пов'язані з мовою, склад і вживання, загальні закономірності пристосування та розвиток мови;

- прикладна лінгвістика спеціалізується на вирішенні практичних завдань, пов'язаних з вивченням мови, а також на практичному використанні лінгвістичної теорії в інших областях. Наприклад методика вивчення рідної та нерідної мови, лексикографія,

переклад, дешифрування, орфографія, транслітерація, розробка термінології і тому подібне;

- практична лінгвістика є сферою, де реально проводяться лінгвістичні експерименти, що мають на меті верифікацію положень теоретичної лінгвістики та перевірку ефективності продуктів, що створюються прикладною лінгвістикою. Наприклад навчання дітей рідній мові, вивчення іноземної мови, переклад, викладання рідної та іноземної мови, літературне редагування, коректура, практична логопедія, побутова і художня словотворчість, мовна політика, створення нових писемностей і навчання грамоті, і тому подібне) [2].

В прикладній лінгвістиці можна виділити наступні напрямки, пов'язані з вивченням мови:

- лексикографія – теорія та практика складання словників;
- лінгводидактика – наука про розробки методик навчання іноземній мові;
- термінознавство – наука про впорядкування та стандартизацію науково-технічної термінології;
- перекладознавство – теорія перекладу.

Основними напрямками прикладної лінгвістики (англ. applied linguistics), які пов'язані з практичними впровадженнями, є наступні:

- Комп'ютерна лінгвістика (англ. computational linguistics):
 - корпусна лінгвістика (створення та використання електронних корпусів текстів);
 - машинний переклад;
 - розробка систем пошуку та класифікації;
 - автоматичне розпізнавання символів (системи автоматичної корекції правопису, такі як автокоректори, спелчекери і так далі);
 - автоматичне реферування текстів (системи авто реферування, наприклад Automatic Text Summarization);
 - створення електронних словників, тезаурусів, онтологій (електронні одномовні та перекладні словники загального вживання: тлумачні,

орфографічні, термінологічні, тощо, наприклад LINGVO фірми АBBYY; КОНТЕКСТ фірми ІН-ФОРМАТИК; УЛІС фірми ПроЛінг);

– автоматичне розпізнавання мови (системи акустичного розпізнавання – Automatic Speech Recognition, і синтезу мовлення – Text-To-Speech Engine);

– оптичне розпізнавання символів (системи оптичного розпізнавання, наприклад OCR – Optical Character Recognition, система OCR Fine Reader фірми АBBYY);

– логічний аналіз текстів (системи логічного аналізу змісту текстів та локалізації знань, наприклад впровадження фірмою Text Analysis International Inc. наприкінці 2001 року інтегровано системи розробки Visual Text™ для побудови програмного забезпечення глибокого аналізу тексту).

• Лінгвістична експертиза. Один з видів лінгвістичного дослідження, який призначається уповноваженою особою (органом) з метою встановлення юридично значущих фактів.

У лінгвістичному аспекті лінгвістична експертиза – це вид дослідження об'єктів, що встановлює істинність / хибність або можливість / неможливість висловлювань опису об'єкта (об'єктів). Об'єктом лінгвістичної експертизи є продукти мовленнєвої діяльності (висловлювання, тексти, лексеми, словесні позначення товарних знаків, тощо). Предметом експертизи є певний лінгвістичний зріз досліджуваного об'єкта.

Основні задачі, які вирішує лінгвістична експертиза наступні:

– виявлення змісту слів, фраз, пропозицій, що містяться у суперечливих текстах (на електронних носіях та в мережі Інтернет, офіційних і приватних документах, тощо) і встановлення форми їх вираження; встановлення змісту поняття, що виражається словом, словосполученням, пропозицією; встановлення ступеня адекватності інтерпретації одного тексту іншим текстом (визначення змісту) та інше;

– оцінка тексту з точки зору вирішення питання про його автора; визначення авторських, суперечливих, анонімних текстів, різновиди літературного плагіату та інше;

– встановлення схожості товарних знаків; встановлення наявності в найменуваннях прихованої та явної пропаганди (наркотиків, насильства, міжнародної ворожнечі, тощо), прихованої реклами і антиреклами та інше.

• Наука про впорядкування та стандартизацію науково-технічної термінології. Об'єктом упорядкування у термінознавстві є термінологія, тобто природно сформована сукупність термінів певної галузі знання або її фрагменту. Термінологія піддається систематизації, потім аналізу з метою виявлення її недоліків та методів їх усунення і, нарешті,

нормалізації. Результат цієї роботи представляється у вигляді терміносистеми – упорядкованої множини термінів із зафіксованими відносинами між ними, що відображають відносини між поняттями, що називають ці терміни.

Таким чином, задача систематизації знань про лінгвістику як науку для спрощення розуміння етапів обробки лінгвістичної інформації та її подальшої реалізації за допомогою обчислювальної техніки є актуальною на сьогоднішній день.

Метою даної роботи є представлення комп'ютерної лінгвістики у вигляді ієрархічної структури та побудова функціональних моделей її елементів.

1. Побудова структури комп'ютерної лінгвістики

Комп'ютерна лінгвістика – це напрямок у прикладній лінгвістиці, який вивчає застосування математичних моделей для опису лінгвістичних закономірностей. За способом дослідження лінгвістичної інформації комп'ютерна лінгвістика поділяється на дві великі частини:

• вивчення способів застосування обчислювальної техніки в лінгвістичних дослідженнях (застосування відомих математичних методів, наприклад статистичної обробки, для виявлення лінгвістичних закономірностей);

• осмислення текстів, написаних природною мовою (створення математичних моделей для розв'язання лінгвістичних завдань та розробка програм, які функціонують на основі цих моделей). Друга частина комп'ютерної лінгвістики тісно пов'язана з розділом штучного інтелекту, що займається розробкою систем опрацювання природної мови (йдеться про засоби обробки як текстової інформації, так і природного мовлення).

У вузькому сенсі проблематику комп'ютерної лінгвістики часто пов'язують з міждисциплінарним прикладним напрямком «обробка природної мови», тобто розробка методів, технологій і конкретних систем, які забезпечують спілкування людини з ЕОМ природною або обмежено природною мовою.

На рис. 1 схематично представлена структура комп'ютерної лінгвістики. Далі розглянемо більш детально кожний з підрозділів комп'ютерної лінгвістики.

2. Корпусна лінгвістика

Корпусна лінгвістика – розділ комп'ютерної лінгвістики, у якому дослідження проводяться за допомогою комп'ютерних лінгвістичних корпусів. Під лінгвістичним корпусом текстів розуміємо великий, уніфікований, структурований, розмічений, філологічно компетентний масив мовних даних, представлений в електронному вигляді й призначений для розв'язання різних лінгвістичних завдань [3].

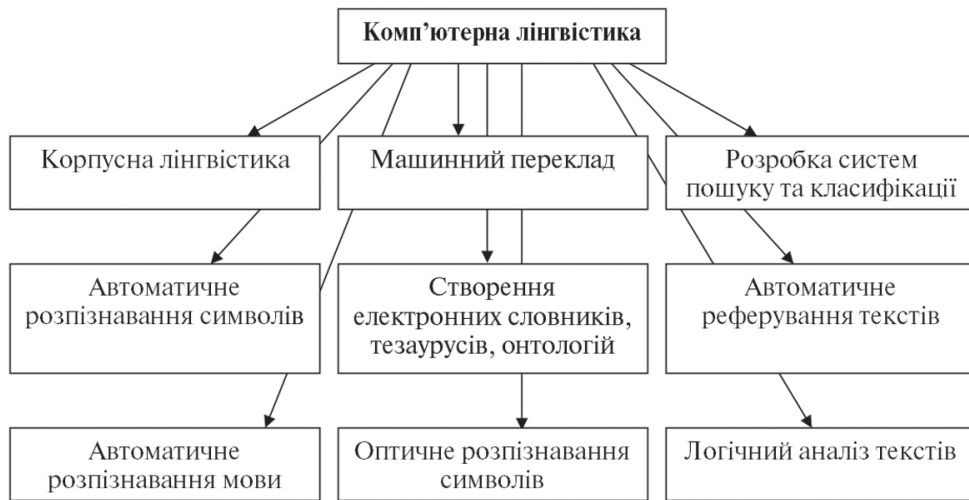


Рис. 1. Структура задач комп'ютерної лінгвістики

Першим великим комп'ютерним корпусом вважається Брауновський корпус, створений у 1960-х роках в Університеті Брауна, що містив 500 фрагментів текстів по 2 тисячі слів у кожному, які були опубліковані англійською мовою в США в 1961 році. У результаті він став стандартом в 1 млн. слововживань для створення представницьких корпусів на інших мовах. У зв'язку із зростанням комп'ютерних потужностей, здатних працювати з великими обсягами текстів, у 1980-і роки в світі було зроблено кілька спроб створити корпуси більшого розміру. У Великобританії такими проектами були Банк Англійської Мови (Bank of English) і Британський Національний Корпус (British National Corpus, BNC). В Росії таким проектом був Машинний Фонд російської мови, який було створено за ініціативою А. П. Єршова. В Україні таким проектом є репрезентативний лінгвістичний кор-

пус Українського мовно-інформаційного фонду НАН України, який функціонує в промисловому режимі приблизно з 2000 року і налічує понад 25 млн. слововживань [3, 4].

Найважливішим при створенні лінгвістичних корпусів є поняття розмітки, відмежовування корпусів текстів від простого зібрання електронних текстів. Розмітка полягає в тому, що текстам корпусу та його компонентам приписуються спеціальні мітки (індикатори) різних типів:

- зовнішні (зазвичай, це елементи бібліографічного опису: видання, рік, автор, тощо);
- структурні (описують структуру тексту: розділ, абзац, речення, тощо);
- лінгвістичні (лексикографічні та граматичні характеристики).

На рис. 2 представлено функціональну модель корпусної лінгвістики.

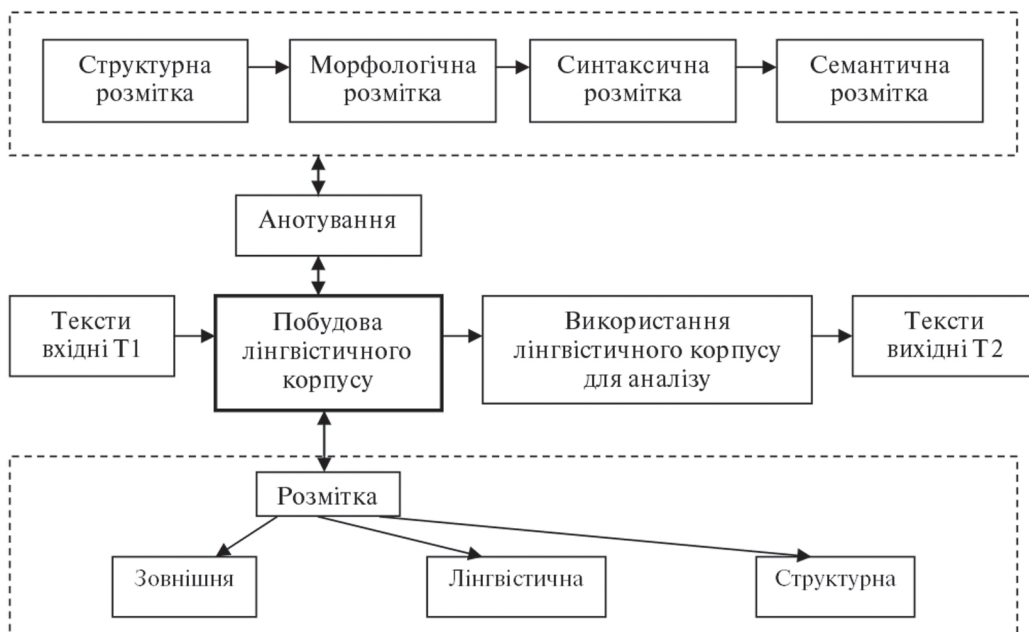


Рис. 2. Функціональна модель побудови лінгвістичних корпусів

3. Машинний переклад

Під час машинного перекладу (МП) можна зустрітись з такими проблемами:

- проблема створення великих словників для систем;
- проблема навчання системи розпізнавання стійких виразів;
- проблема запису усіх правил побудови речення і перекладу у вигляді програми.

Одна із систем, яка намагається вирішити ці проблеми, розробляється в Українському мовно-інформаційному фонді. В цій системі використовуються принципи індуктивно-інтерактивного моделювання процесу автоматичного перекладу з незалежними етапами аналізу вхідного тексту й синтезу вихідних текстів [5].

На рис. 3 представлена схема процесів машинного перекладу.

На першому етапі машинного перекладу відбувається аналіз вхідного тексту, який складається з чотирьох модулів аналізу:

- передморфологічний;
- морфологічний;
- контекстний;
- синтаксичний.

Кожен з цих модулів може використовуватися також і в інших системах обробки мовної інформації (автоматичне індексування, інформаційний пошук, автоматичне редагування, тощо). Сама ж

система аналізу є основою лінгвістичного забезпечення формального опису мови.

Алгоритми перекладу будуються на інформації про міжмовні відношення щодо лексичних, морфологічних та семантико-синтаксичних відповідностей між одиницями перекладу вхідного тексту і синтезованого вихідного. Регульовальну роль у цьому процесі відіграє граматики, що фіксує синтаксичні нормативні правила вихідної мови. Трансформаційні граматики (знаходження подібності та наявність у синтаксичній системі вихідних структур і правил їх перетворення (трансформації) за умови збереження незмінним лексичного складу вихідного змісту і синтаксичних відносин між лексемами) будуються на зіставленні грамастик обох мов.

Ядро системи машинного перекладу становить автоматичний багатомовний перекладний словник, який є компонентом інтегрованої лексикографічної бази Українського мовно-інформаційного фонду. Автоматичний багатомовний перекладний словник будується як відкрита неоднорідна, багатофункціональна лексикографічна база даних [6]. Словник інтегрує в собі лексикони мов, залучених до системи МП, забезпечує пошук перекладних еквівалентів з української мови на інші і навпаки. Автоматичний багатомовний перекладний словник формується на основі двомовних електронних словників із залученням комп'ютерних термінологічних баз даних, електронних тлумачних словників, словників синонімів, синтаксичних словників, що фіксують валентні властивості лексичних і граматичних мовних одиниць. Усі ці словники заносяться до загальної лексикографічної бази.

Автоматичний багатомовний перекладний словник інтегрується з граматичними словниками відповідних мов.

Граматичні словники використовуються як базовий інструмент на всіх етапах аналізу і синтезу МП.

Крім автоматичного багатомовного перекладного словника та граматичних словників до словникової бази включаються бінарні перекладні словники словосполучень як основний засіб представлення контекстного визначення багатозначних слів.

Для словників словосполучень використовується лінгвістична база з корпусів паралельних текстів для кожної пари мов, на які орієнтована система МП. Ця ж база є вихідним мовним матеріалом для розробки лінгвістичних алгоритмів аналізу, синтезу і трансформаційних грамастик, що лежать в основі алгоритмів перекладу.

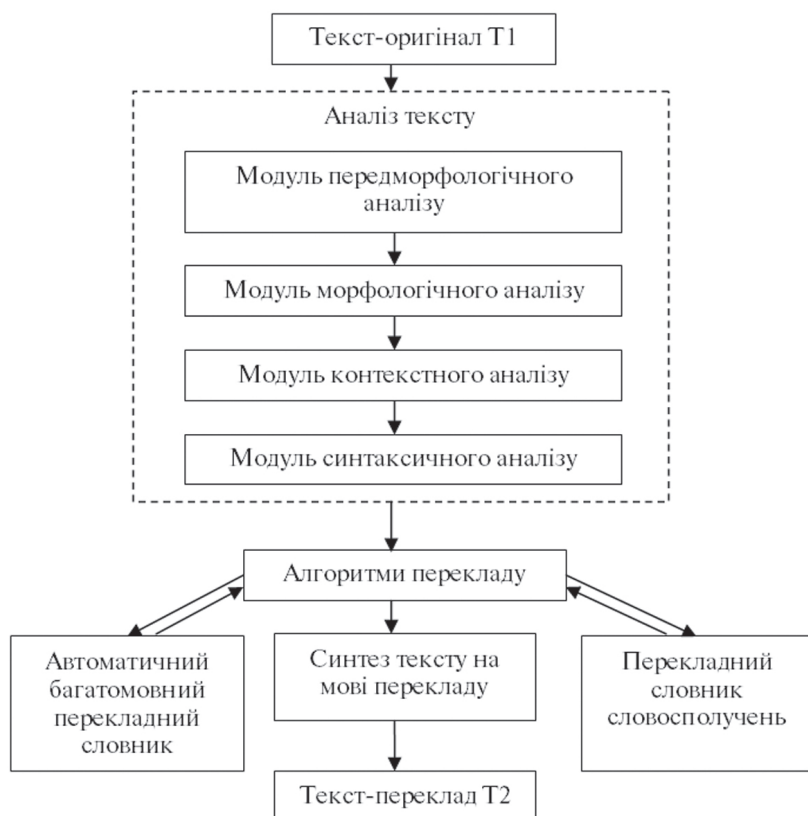


Рис. 3. Функціональна модель машинного перекладу

4. Системи пошуку та класифікації

Інформаційно-пошукова система виконує такі функції:

- зберігання великих обсягів інформації;
- швидкий пошук необхідної інформації;
- внесення, видалення і зміни інформації, що зберігається;
- подання вихідної інформації в зручному для людини вигляді.

На рис. 4 представлена загальна схема етапів пошуку та класифікації інформації.

Велике значення в сучасних повнотекстових інформаційно-пошукових системах приділяється морфологічному аналізу, тобто автоматичним засобам обробки окремих слів (як в текстах вихідних документів, так і в запитах користувача). Хоча слід зауважити, що не всі пошукові системи включають до своєї системи морфологічну обробку тексту (як приклад можна назвати систему AltaVista (<http://www.altavista.com/>). Недоліком цієї системи є слабе ранжування (ранжування визначає порядок видачі результатів пошуковим сервером: так, сторінки, на яких слова запиту зустрічаються частіше, отримують більш високий ранг і виводяться першими).

Використовуючи морфологічний аналіз в пошукових системах, розробники ставлять за мету покращення результату пошуку, а саме: збільшення повноти та точності пошуку. Коли відбувається пошук за канонічними формами слів, до результату пошуку потрапляють не тільки документи зі словом у тій формі, яка точно відповідає словоформі у запиті, але й інші документи, які містять різні форми даного слова (тобто обсяг інформації, що видається на запит користувача, справді збільшується). Ще однією перевагою використання канонічних індексних баз під час пошуку є спрощення інтерфейсу користувача. Найчастіше користувачу необхідно отримати на запит всі варіанти відповідей на запит. За відсутності автоматичного розширення слів його варіантами (коли пошук виконується не за однією, вказаною у запиті, словоформою, а за множиною усіх форм даного слова, що входять до його словозмінної парадигми) користувач зму-

шений вивчати і використовувати у своїх запитах формули або оператори відсікання.

Багато праць присвячено дослідженню впливу різних алгоритмів морфологічного аналізу на якість інформаційного пошуку в текстах на різних мовах. Результати цих досліджень дозволяють зробити висновок, що запровадження морфологічних модулів у пошукові системи дозволяє збільшити повноту і точність інформаційного пошуку [7]. Яскравим прикладом таких систем є програмні продукти серії Yandex (Мовний index).

Висновки

Нами було запропоновано та розроблено представлення комп'ютерної лінгвістики у вигляді ієрархічної структури та структурний опис задачі побудови функціональних моделей обробки мовної інформації на відповідних етапах.

Крім розглянутих вище підрозділів комп'ютерної лінгвістики, таких як корпусна лінгвістика, машинний переклад і системи пошуку та класифікації, є перспективи подальших досліджень процесів обробки інформації у таких напрямках, як:

- Автоматичне розпізнавання символів (системи автоматичної корекції правопису (автокоректори, спелчекери). Наприклад: ОРФО – система перевірки правопису, розробка фірми ІНФОРМАТИК; РУТА – система перевірки правопису; УЛІС – електронний українсько-російський та російсько-український словник; Ispell – GNU Project – Free Software Foundation, URL.
- Автоматичне реферування текстів (системи авто реферування. Наприклад: Inxight – комерційна система автоматичного реферування; Extractor – Модуль авто реферування; Text Analyst™ – програмний інструмент для аналізу змісту текстів, змістовного пошуку інформації, формування електронних архівів.
- Створення електронних словників, тезаурусів, онтологій (електронні одномовні та перекладні словники загального вжитку (глумачні, орфографічні, термінологічні тощо)). Наприклад: LINGVO фірми АВВУ, КОНТЕКСТ фірми ІНФОРМАТИК, УЛІС фірми ПроЛінг.

«АВВУ Lingvo 11 Багатомовна версія» – найбільш популярний багатомовний електронний словник, що містить 113 професійних словників і дозволяє перекладати з російської мови на іспанську, англійську, німецьку, французьку та італійську мови. Крім традиційного списку європейських мов, перекладач дає можливість працювати з нещодавно доданими китайською, турецькою, українською, португальською мовами, а також робить доступним словник афоризмів латинської мови.

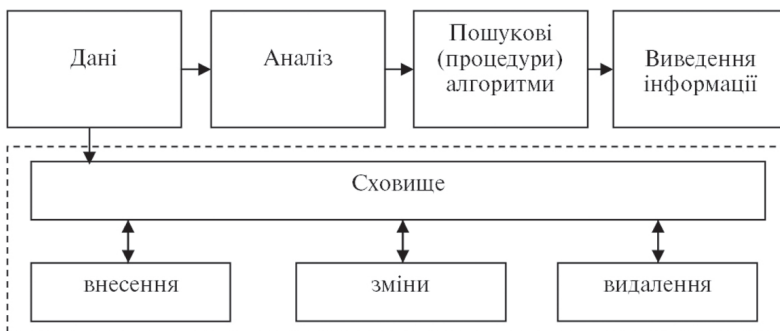


Рис. 4. Функціональна модель системи пошуку та класифікації

Електронні словники КОНТЕКСТ – це система професійних електронних словників. Містить великий набір багатомовних словникових баз як загальнолексичних, так і диференційованих по різних галузях знань. Активний розвиток електронного словника КОНТЕКСТ дозволило йому стати зручним і надійним помічником для тих, хто вивчає іноземні мови або працює з ними.

УЛІС – електронний російсько-український і українсько-російський словник. Основні можливості:

– в словникові статті включені докладні тлумачення значень слів, граматичні коментарі й приклади вживання слів у сталих словосполученнях;

– для багатозначних слів широко представлені варіанти перекладу;

– зручна пошукова система дозволяє легко і швидко знайти потрібне слово в будь-якій формі, його синоніми або близькі за значенням слова;

– додатково в словнику УЛІС реалізована можливість перекладу слова з будь-якої його форми з урахуванням особливостей морфології.

• Автоматичне розпізнавання мови (системи акустичного розпізнавання) і синтезу мовлення. Наприклад, <http://www.speech.com.ua> – сайт з розпізнавання та синтезу мовлення в Україні; Voice Type Dictation, Voice Pilot и ViaVoice от IBM : <http://www.ibm.com/software/speech>.

• Оптичне розпізнавання символів (системи оптичного розпізнавання). Наприклад система OCR Fine Reader фірми ABBYY.

Більшість програм оптичного розпізнавання тексту (OCR Optical Character Recognition) працюють з растровим зображенням, яке отримано через факс-модем, сканер, цифрову фотокамеру або інший пристрій.

Система ABBYY FineReader призначена для конвертації в редаговані формати відсканованих документів, PDF-документів та файлів зображень, включаючи цифрові фотографії.

Список літератури: 1. Четвериков, Г.Г. Концептуально-методологічний підхід до моделювання природної мови алгебро-логічними засобами [Текст] / І.Д. Вечірська, Т.Н. Федорова та ін. // Тезиси доклади Международной

научной конференции “Горизонты прикладной лингвистики и лингвистических технологий” (MegaLing’2009). – 20-27 сентября 2009, Украина, Киев. – С. 68. 2. *Лингвистика* [Электронный ресурс] / Режим доступа : [www/URL: http://ru.wikipedia.org/w/](http://ru.wikipedia.org/w/) – 24.02.2010 г. – Загл. с экрана. 3. Широков, В.А. Корпусна лінгвістика [Текст] / В.А. Широков, О.В. Бугаков, Т.О. Грязнухіна та ін. ; відповідальний редактор В.А. Широков. – К. : „Довіра”, 2005. – 472 с. 4. *Корпусная лингвистика* [Электронный ресурс] / Режим доступа : [www/URL: http://ru.wikipedia.org/wiki/](http://ru.wikipedia.org/wiki/) – 11.02.2010 г. – Загл. с экрана. 5. Широков, В.А. Очерк основных принципов квантовой лингвистики [Текст] / В. А. Широков // Бионика интеллекта. – 2007. – № 1 (66) – С. 25 – 32. 6. Грязнухіна, Т.О. Автоматичний багатомовний перекладний словник [Текст] / Т.О. Грязнухіна, Т.П. Любченко – Вісник лінгвістичного університету, том 6. – № 2. – Київ, 2003. – С. 68-71. (Особистий внесок: організація даних та структура АБПС). 7. Губин, М.В. Влияние морфологического анализа на качество информационного поиска [Текст] / М.В. Губин, А.Б. Морозов : Труды Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL2006 URL: http://www.rcdl2006.uniyar.ac.ru/papers/paper_67_v2.pdf.

Надійшла до редколегії 25.03.2010 р.

УДК 519.62

Построение функциональных моделей элементов компьютерной лингвистики / И.Д. Вечирская, Т.Н. Федорова, Г.Г. Четвериков // Бионика интеллекта: науч.-техн. журнал. – 2010. – №1 (72). – С. 83–88.

В статье было предложено и разработано представление компьютерной лингвистики в виде иерархической структуры и структурное описание задачи построения функциональных моделей обработки языковой информации на соответствующих этапах.

Л. 4. Библиогр.: 7 назв.

UDK 519.62

Construction functional models of computer linguistics elements / I.D. Vechirska, T.N. Fedorova, G.G. Chetverikov // Bionics of Intelligence: Sci. Mag. – 2010. – № 1 (72). – P. 83–88.

In article it was offered and representation the computer linguistics in the form of hierarchical structure and the structural description problem of construction functional models processing the language information at corresponding stages is developed.

Fig. 4. Ref.: 7 items.