УДК 811.93.004.82

Г. О. Хацер

Дніпропетровський державний університет внутрішніх справ

НОВІТНІ ТЕНДЕНЦІЇ У ЛІНГВІСТИЧНІЙ ОБРОБЦІ ТЕКСТУ ПРОГРАМНИМИ ЗАСОБАМИ

Розглянуто проблему лінгвістичної обробки тексту. Проаналізовано новітні тенденції у розвитку методів та шляхів дослідження тексту за допомогою автоматичних засобів обробки. Узагальнено поняття «текст». Описано рівні та компоненти автоматичної обробки текстів. Розкрито зв'язок лінгвістики тексту з комп'ютерною лінгвістикою. Визначено, що при аналізі тексту комп'ютерними засобами відбувається чіткий розподіл функцій.

Ключові слова: обробка тексту, автоматичні засоби обробки текстів, текст, лінгвістика тексту, комп'ютерна лінгвістика.

Хацер А. А. Днепропетровский государственный университет внутренних дел.

НОВЕЙШИЕ ТЕНДЕНЦИИ В ЛИНГВИСТИЧЕСКОЙ ОБРАБОТКЕ ТЕКСТА ПРОГРАММНЫМИ СРЕДСТВАМИ

Рассмотрена проблема лингвистической обработки текста. Проанализированы новейшие тенденции в развитии методов и путей исследования текста при помощи автоматических средств обработки. Обобщено понятие «текст». Описаны разные уровни и компоненты автоматической обработки текстов. Раскрыта связь лингвистики текста с компьютерной лингвистикой. Определено, что при анализе текста компьютерными средствами происходит четкое разделение функций.

Ключевые слова: обработка текста, автоматические средства обработки текстов, текст, лингвистика текста, компьютерная лингвистика.

Khatser G. O. Dnipropetrovsk State University of Internal Affairs.

MODERN TENDENCIES IN LINGUISTIC PROCESSING OF THE TEXT BY SOFTWARE TOOLS

The object is the existing trends and methods of linguistic processing of the text by software tools. The object of the article is linguistic processing of texts by software tools. The aim of the article is to analyze and to generalize data on existent tendencies in linguistic processing of the text by software tools. The following tasks are to be fulfilled to reach the goal set: 1) to study the current trends in linguistic processing of the text; 2) to single out main characteristics and features of linguistic computer processing of texts; 3) to generalize and systematize the existent data on linguistic processing of the text by software tools.

While analyzing the text by computer means, the strict division of functions is important. The final product of the analysis is certain semantic structure. Linguistic knowledge such as morphological and syntactic analysis is the ground of developing of programs on linguistic texts processing. The work of analyzers is built on the existent body of lexical units, dictionaries, databases and so on. In modern terms of the world's development there is a tendency of moving from simple syntactic and morphological analysis to semantic one of the whole text. Thus, it is appropriate to continue scientific research in the field of peculiarities of semantic analysis of texts.

Key words: processing of the text, software tools of text's processing, text, linguistics of the text, computer linguistics.

На сьогоднішній час філологічна проблема лінгвістичного розбору текстів вийшла за межі тільки однієї науки. Цим питанням займаються також провідні фахівці в галузі програмних розробок. Поява вагомого електронного банку текстових ресурсів значно вплинула на подальший розвиток лінгвістичної обробки тексту автоматичними засобами й вивела задачу на новий рівень.

Об'єкт дослідження – існуючі направлення та методи лінгвістичної обробки тексту за допомогою програмних засобів. Предмет статті – лінгвістична обробка текстів програмними засобами.

Мета статті – проаналізувати та узагальнити дані про існуючі тенденції у лінгвістичній обробці тексту автоматизованими засобами. Для досягнення поставленої мети необхідно реалізувати такі ***завдання***: 1) проаналізувати існуючі тенденції лінгвістичної обробки текстів; 2) виявити основні ознаки та характеристики лінгвістичної комп'ютерної обробки текстів; 3) узагальнити та систематизувати існуючі дані стосовно лінгвістичної обробки тексту програмними засобами.

З філологічного погляду лінгвістика тексту – це одне з направлень досліджень, об'єктом яких виступають: 1) правила, на базі яких відбувається побудова цільного та зв'язаного тексту; 2) смислові категорії, що відображаються у відповідності до цих правил [2, с. 15]. Її самостійне ставлення відбулося у 60-ті роки XX століття у Німеччині. Вивчення відбувається за декількома напрямками:

– структурно-граматичне: головна увага приділяється формальним засобам та типам зв'язаності тексту, принципам побудови структури тексту;

– семантичне: виявлення відношень всередині самого тексту;

– комунікативно-прагматичне: дослідження функцій тексту в системі дискурсу як знакового посередника між автором та вчителем, встановлення прагматичних інтенцій та стратегій тексту;

– семіотичне: відношення текстового знака, його денотата та реальності, рекурсивні та прокурсивні зв'язки тексту з іншими текстами та продуктами культури з погляду категорій модальності та інтертекстуальності;

– когнітивне: дослідження змісту тексту шляхом моделювання когнітивних структур репрезентацій знань, які призводять до появи та розуміння самого тексту;

– прикладне: інтегрує лінгвістику тексту з комп'ютерною галуззю мовознавства, оскільки моделі когнітивного сприйняття застосовуються при автоматичній обробці природної мови, формуванні систем синтезу та аналізу текстів, багаторівневих лінгвістичних процесів, а також у машинному перекладі [8, с. 333].

Для подальшої роботи необхідно виділити три найбільш значимих текстових рівні:

1) структурно-семантичний, якого дотримуються А. Богуславський та Ю. О. Сорокін. Учені вважають головною характеристикою тексту структурну організацію;

2) комунікативний, представниками якого вважаються О. О. Реформатський та Р. Барт. Особливість підходу – структурно-семантична організація співвідноситься з людським фактором та комунікативними завданнями;

3) психолінгвістичний – Т. М. Дрідзе співвідносить текст із процесами людської діяльності та визначає його як головну одиницю спілкування [4, с. 37].

Для автоматичної обробки тексту необхідно враховувати його рівні та компоненти, серед яких смисл, ступінь інформованості та стильова специфіка (див. далі рис. 1).

Крім цього, існує авторський рівень, який реалізує задум автора через комунікативні програми, розкриваючи теми та підтеми. Така складна структура тексту визиває низку труднощів при обробці та відтворенні тексту комп'ютерними засобами.

У статті за головні визначення та розуміння тексту приймаються наступні:

1) текст як галузь функціонування мови, його використання у мовленні;

2) текст як вищий рівень мовної системи – одиниці мови;

3) текст як одиниця спілкування, яка має певну змістовну завершеність і кінцевим результатом якої виступає реалізація комунікативних цілей.

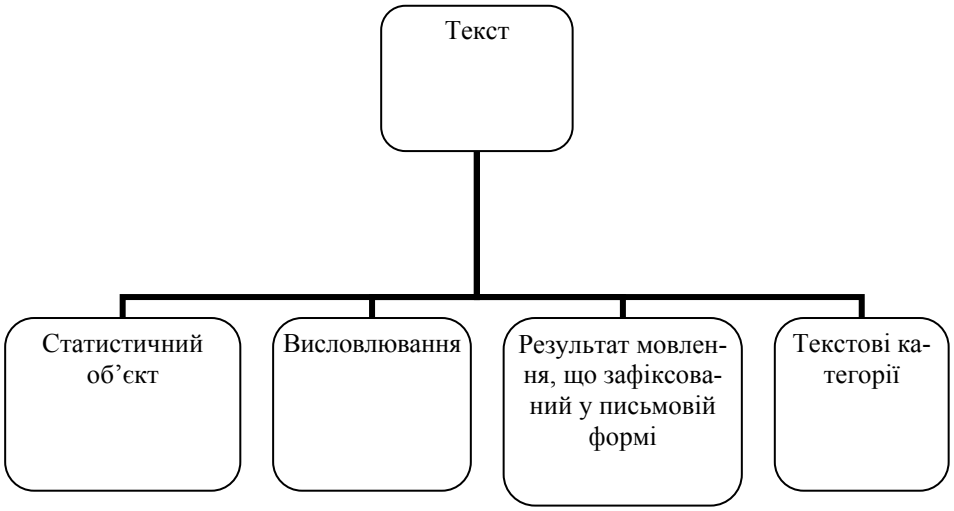

Рис. 1. Складові елементи розуміння поняття «текст»

Робота з текстами неможлива без розуміння терміна «мова», оскільки саме мова є засобом передачі інформації, що міститься в самому тексті.

Обробкою текстів у межах комп'ютерних розробок займаються вітчизняні й закордоні вчені, серед яких Є. Р. Айвазова [1], Є. І. Большакова, Е. С. Клишинський [2], Н. М. Леонтьєва [6], R. Delmonte [9], A. Gelbukh [10], J. Geiß [11], S. Toru [12] та ін.

Н. М. Леонтьєва, одна з провідних спеціалістів у галузі прикладного підходу до трактування тексту, будує свої дослідження на «м'якому» розумінні тексту, яке зводиться до підлаштовування роботи автомата у відповідності з різними комунікативними завданнями. При аналізі тексту комп'ютерними засобами відбувається чіткий розподіл функцій: людина встановлює ціль, а завдання машини – зрозуміти кінцевий продукт (текст), що був заданий людиною, та оцінити результати на всіх рівнях. Кінцевий продукт – певна семантична структура. Автор виокремлює наступні рівні розуміння тексту:

– локальний (лінгвістична структура речень): розуміння тексту, яке обмежене рамками речення. Основа такого представлення – синтаксичне дерево речення з семантичними вузлами / зв'язками. Робота програми зводиться до побудови складних дерев на базі словникових статей. Якщо у словнику немає необхідної статті, то побудова такого дерева стає неможливою. А це вважається одним із недоліків даного підходу. Інший недолік – це неможливість вийти за межі речення, а смисл тексту – це не сума смислів усіх речень, що входять до його структури.

– глобальний (семантична структура): трансформація усіх семантико-синтаксичних структур речень тексту на більш елементарні одиниці. Головний акцент припадає на комунікативні відношення як всередині самого речення, так і між самими реченнями.

– спеціальний або вибірковий (структура баз даних та знань): спеціальне розуміння, яке в найбільш повній формі ураховує екстралінгвальні аспекти тексту, відображаючи певну частину реалій. На цьому рівні розуміння оперує такими поняттями, як структура баз даних та структура баз знань. Під першою розуміють формальні, фіксовані структури, над якими можливо здійснювати математичні дії [6, с. 52]. Під другою – динамічні структури, що відображають цілий текст, а не

результат членування речення. Відповідно до Р. Шенко, їх головна задача – впізнавання певного сюжету всередині тексту.

При автоматичній обробці тексту та створенні програмних продуктів постає питання про урахування усіх рівнів текстового розуміння, що були окреслені вище. Це породжує труднощі, що пов'язані з багатозадачністю та багаторівневістю поставленої мети. Так, при виокремленні відношень у тексті необхідно задіяти графовий та морфологічний аналіз, виділити конструкції на семантичному рівні. Існує низка лінгвістичних технологій, систем, утиліт для лінгвістичного аналізу та обробки текстів. Головні розробки здійснюються вченими у Сполучених Штатах Америки та Росії.

«Інтел Сервіс» розробив семантичні системи AskNet, що будуються на принципі «питання – відповідь», та реалізують повний лінгвістичний аналіз російськомовних та англомовних текстів. Російська компанія RCO розроблює низку програм для морфологічного, синтаксичного аналізу текстів російською мовою. Взагалі лінгвістичний аналіз включає словники графем, морфологічні словники, словники синонімів, правил синтаксису та інші. Використовується морфологічний словник приблизно на 115000 лексем.

Програма Galaktika-Zoom – це автоматизована система пошуку та обробки інформації. Вона дозволяє виокремити значимі слова та словосполучення текстових документів та проводить пошук за ключовими словами, що були задані користувачем; отримувати якісний інформаційний результат у короткі строки. При здійсненні запиту програма додатково створює список слів та словосполучень, які вважаються важливими для цього запиту. Крім цього, існує можливість укладати звіти про частоту попадання слів у російськомовних документах.

Більшість програм було розроблено для реалізації окремих етапів лінгвістичного аналізу текстів. Link Grammar Parser, Mystem, Russian Link Grammar Parser и Apache OpenNLP розроблені для паркінгу російськомовних та англомовних текстів. Програма морфологічного аналізу текстів «морфологічний аналізатор», що була розроблена російським вченим С. А. Старостіним, дозволяє проводити онлайн аналіз англійських та російських слів, незалежно від їх граматичної форми. Основою програми стали словники О. О. Зализняка та В. К. Мюллера [5; 7]. Результат роботи може бути представлений у декількох варіантах:

– словникова стаття зі словника, що входить до бази;

– попередня морфологічна форма;

– переклад слова у відповідності з даними словника В. К. Мюллера;

– цілісна морфологічна характеристика слова, що було введене.

Крім цього, розповсюдженими є програми для математичного аналізу текстів (WordStat, Лінгвоаналізатор), дослідження поведінки слів у тексті WordSmith Tools), стилістичної перевірки російськомовних текстів (Fresh Eye) та інші.

Таким чином, лінгвістичні знання, а саме основи морфологічного та синтаксичного аналізу, слугують базою для створення програм лінгвістичної обробки текстів. Робота аналізаторів будується на вже існуючому корпусі лексичних одиниць, словників та баз даних.

У сучасних умовах розвитку світового простору існує тенденція до переходу від простого синтаксично-морфологічного аналізу до семантичного аналізу цілого тексту. Тому у подальшому доцільним є дослідження питань особливостей побудови саме семантичного аналізу текстів.

Бібліографічні посилання

1. **Айвазова Э. Р**. Метод семантического анализа лексических значений / Э. Р. Айвазова // Культура народов Причерноморья. – 2007. – № 121. – С. 151–153.
2. **Бабенко Л. Г.** Лингвистический анализ художественного текста / Л. Г. Бабенко, И. Е. Васильев, Ю. В. Казарин. – Екатеринбург, 2000. – 125 с.
3. **Большакова Е. И.** Автоматическая обработка текстов на естественном языке и компьютерная / Е. И. Большакова, Е. С. Клишинский. – М. : МИЭМ, 2011. – 272 с.
4. **Дридзе Т. М.** Текстовая деятельность в структуре социальной коммуникации. Проблемы семиосоциопсихологии / Т. М. Дридзе. – М. : Наука, 1984. – 232 с.
5. **Зализняк А. А.** Грамматический словарь русского языка. Словоизменение / А. А. Зализняк. – М. : Рус. язык, 1980. – 880 с.
6. **Леонтьева Н. Н.** Автоматическое понимание текста: системы, модели, ресурсы / Н. Н. Леонтьева. – М. : ИЦ «Академия», 2006. – 98 с.
7. **Мюллер В. К.** Англо-русский словарь / В. К. Мюллер. – М. : Рус. язык, 1995. – 2106 с.
8. **Селіванова О. О.** Лінгвістична енциклопедія / О. О. Селіванова. – Полтава : Довкілля-К, 2010. – 844 с.
9. **Delmonte R**. Computational Linguistic Text Processing / R. Delmonte. – Texas, 2010. – 382 p.
10. Computational Linguistics and Intelligent Text Processing // 5th International Conference, CICLing 2004, Seoul, Korea, February 15–21, 2004, Proceedings Series: Lecture Notes in Computer Science, Vol. 2945 / Editors: A. Gelbukh. – 2004. – 658 p.
11. **Geiß J.** Latent semantic sentence clustering for multi-document summarization / J. Geiß. – Cambridge, 2011. – 156 p.
12. **Toru S.** A Computational Framework for Text Processing Based on Systemic Functional Linguistics [Electronic resource] / S. Toru. – Access mode : http://www.brain.riken.jp/labs/mns/sugimoto/csfgc05.pdf.

Надійшла до редколегії 24.02.15

УДК 821.111

V. V. Yashkina

Oles Honchar Dnipropetrovsk National University

POETRY TRANSLATION ADEQUACY: MULTILINGUAL EXPERIENCE OF HEANEY'S TRANSLATIONS

The problems of a poetic text translation have been of current relevance since the appearance of translations as they are, but have not received final solution. Linguists are proving impossibilities of full conceptual, semantic and language preservation of the original in translation. Professionals and amateurs never stop arguing on the dilemma of equivalence versus adequacy. Translators offer numerous variants of one and the same poetic masterpiece with the aim of achieving maximum original beauty and sense. As a result, it becomes obvious, that the more different language translation samples appear, the better the original message of a poem is restored.

In the given article we concentrated our attention on multilingual (Spanish, Ukrainian, Russian) translations of Seamus Heaney's selected poems, specifically comparing grammar, lexical, phonetic and strophe-building translation variants and discussing discrepancies between the original and its Romanic and Slavic translations. The main aim lies in an attempt to prove that multilingual translation practice serves as a source for translation variants to add to each other or open different aspects of the original, and, possibly, to achieve perfect revealing of a poetic masterpiece.

It has been proved by the research, that noncoincidence of certain translation equivalents accompanied by full correspondence of others, results in definite semantic loss. This fact underlines