УДК81’322

Т. В. Бобкова
Київський національний лінгвістичний університет

КОРПУСНО-БАЗОВАНИЙ ПІДХІД ДО МЕРЕЖЕВОГО МОДЕЛЮВАННЯ ЗНАЧЕННЯ ТЕКСТУ

Реферат. *Мережеве моделювання семантики тексту розглядається в аспекті загальної проблеми автоматичного аналізу тексту й пошуку інформації. У дослідженні здійснено аналіз основних лінгвістичних підходів до мережевого моделювання значення тексту, які ґрунтуються на позамовних класифікаціях предметів зовнішнього світу й диференціюються за застосовуваними методами моделювання. Обґрунтовано необхідність і перспективи використання корпусно-базованого підходу, який сприяє автоматизації й аналізу семантики тексту на підставі поєднання статистичного й корпусного методу.*

У роботі встановлено основні етапи мережевого моделювання семантики тексту на підставі корпусно-базованої методики. Пропонована методика мережевого моделювання ґрунтується на статистичному опрацюванні тексту та аналізі конкордансів і уможливлює отримання й упорядкування об’єктивних даних про семантичні зв’язки між компонентами значення тексту. У перспективі поєднання статистичного й корпусного методів сприятиме побудові просторової тривимірної моделі значення тексту. Застосування колекції текстів більшого обсягу забезпечить підвищення ефективності отриманих результатів.

Ключові слова*: мережеве моделювання, текст, значення, семантичний компонент, корпусно-базований, частота.*

Відомо, що розв’язання більшості прикладних завдань із автоматичного опрацювання інформації вимагає аналізу структури значення цілого тексту. Зокрема, до таких завдань зараховують [3, с. 387] переклад іноземною мовою, загальне розуміння та індексацію за темою тексту, реферування та атрибуцію. Необхідність нагального вирішення питань у галузі автоматичного опрацювання тексту сприяла формуванню в сучасній лінгвістиці різних підходів і методик мережевого моделювання значення тексту [1; 3; 4; 7; 9; 10]. Проблема полягає в тому, що пропоновані методики мережевого моделювання мають уможливити отримання об’єктивних даних про ієрархічну структуру значення тексту. З упровадженням корпусних технологій постає необхідність удосконалення методики мережевого моделювання значення тексту на підставі даних статистичного й корпусного аналізу.

Актуальність нашого дослідження визначається затребуваністю в сучасній лінгвістиці інформації щодо структури значення тексту. **Мета** статті – дослідити корпусно-базований підхід до мережевого моделювання тексту. Досягнення поставленої мети передбачає виконання таких **завдань:** 1) проаналізувати основні методики мережевого моделювання семантики тексту, 2) навести аргументи на користь корпусно-базованої методики й 3) запропонувати методику мережевого моделювання значення тексту на підставі аналізу частоти появи слів у певному контексті.

Мережеве моделювання тексту має на меті встановлення набору семантичних компонентів, поєднаних певними типами відношень. Застосування певної методики мережевого моделювання залежить від інтерпретації структури значення тексту й методів моделювання в межах нелінгвістичного (статистичного) та лінгвістичного (традиційного й корпусно-базованого) підходів. Розглянемо більш детально теоретичні принципи й критерії встановлення семантичного зв’язку за лінгвістичним підходом до мережевого моделювання.

Теоретико-методичні принципи мережевого моделювання тексту закладено працями Е.Ф. Скороходько [7] і Н.П. Дарчук [2; 3; 4]. За традиційним підходом підґрунтям для виявлення семантичних зв’язків між основними одиницями тексту – словами, реченнями й абзацами слугують немовні відношення зовнішнього світу. Стосовно аналізу семантичної

структури тексту це встановлення зв'язків між денотатами слів і денотатами речень, тобто ситуаціями за деривaційним критерієм [2, с. 242–243]. Семантичний зв'язок між словами тексту, який відображає зв'язок між «денотатами зовнішнього світу», називається синтагматичним [2, с. 256–257]. Формально наявність семантичного зв'язку між словами в реченні виявляється за допомогою дистрибутивного аналізу. В основу дистрибутивно-статистичного методу покладено гіпотезу щодо «реляційної залежності між семантичними й синтаксичними властивостями слова, яка дозволяє описувати семантику через синтаксис» [2, с. 241; 10]. Формальними показниками семантичного зв'язку між реченнями тексту є релевантні семантичні компоненти, а також певні слова на позначення причино-наслідкових відношень [2, с. 256–259]. Саме множина взаємопов'язаних релевантних слів і становить семантичну мережу досліджуваного тексту. Традиційно для зображення семантичних зв'язків у лексиці й тексті використовують матрицю або граф [1, с. 145; 2, с. 261], останній вважається оптимальним представленням структури значення. Сказане вище дозволяє стверджувати, що за традиційним лінгвістичним підходом до мережевого моделювання тексту основними критеріями виявлення семантичних зв'язків слугують формальні – синтаксичний і ситуаційний.

Однак для побудови системи семантичного аналізу необхідно задати вихідні дані у вигляді лінгвістичних знань і «знань предметної галузі, до якої належить текст» [4, с. 17]. Головним ресурсом опрацювання змісту тексту вважаються тлумачні словники, використовувані для упорядкування опису [1, с. 144–145; 4, с. 17]. Зокрема, за допомогою ідеографічних словників здійснюється моделювання семантики в парадигматичному аспекті [4, с. 17–18]. Саме тому традиційний підхід слід класифікувати як **системно-орієнтований.** Проблема застосування традиційного підходу до моделювання смислу тексту полягає у необхідності «виходити за межі мови і звертатися до зовнішнього світу, до класифікації предметів, уявлень» [3, с. 387]. Крім того, при опрацюванні великих обсягів текстів постає необхідність автоматизації аналізу на підставі «усталеного універсального методу» [3, с. 387].

Перспективи вирішення окреслених проблем вбачаємо в поєднанні лінгвістичних і статистичних критеріїв, покладених в основу встановлення ієрархії семантичних зв'язків у тексті, що уможливлюється завдяки корпусному підходу до мережевого моделювання тексту. Ідея використання пошукового апарату корпусу для мережевого моделювання є достатньо поширеною в сучасній лінгвістиці, що підтверджується низкою досліджень [9; 10; 11; 12; 13], присвячених проблемам побудови корпусно-базованих семантичних моделей. Теоретичним підґрунтям корпусного підходу до мережевого моделювання тексту є, насамперед, твердження про наявність безпосереднього зв'язку між семантикою слова та частотою його вживання в тексті [6, с. 446–447; 8, с. 137]. У зв'язку з цим корпусний підхід ґрунтується на засадах **текстозорієнтованого**: основним джерелом емпіричних даних слугує текст і відображена в ньому «текстова картина світу», а наукова картина вважається лише окремим виявом текстової [5, с. 184]. Саме тому за корпусним підходом мережеве моделювання передбачає укладання текстозорієнтованих словників – частотного й конкордансів.

При цьому має бути врахована якісна відмінність корпусу як джерела даних [14, p. 207–208], а саме: 1) неможливість ідентифікувати тексти як єдину комунікативну подію, 2) формальний характер корпусних параметрів, 3) фрагментарність і вертикальний напрямок конкордансу. Отже, суть корпусного підходу до мережевого моделювання тексту полягає, насамперед, у застосуванні формально-статистичних критеріїв і прийому конкордансингу для встановлення набору семантичних компонентів. Тобто корпусний підхід слід класифікувати як формально-статистичний саме за застосовуваними методами й прийомами аналізу тексту. Загалом корпусно-базована методика мережевого моделювання тексту передбачає таку послідовність процедур: 1. Укладання частотного словника досліджуваного тексту / текстів. 2. Ранжування результатів статистичного аналізу для визначення текстового концепту й набору компонентів значення тексту. 3. Укладання конкордансу / встановлення

колокацій кожного семантичного компонента. 4. Визначення релевантних зв'язків компонентів значення тексту. 5. Встановлення ієрархії рекурсивних зв'язків між семантичними компонентами. 6. Побудова корпусно-базованої семантичної моделі тексту.

У запропонованому дослідженні для побудови корпусно-базованої моделі значення тексту обрано статтю Дж. Лакоффа «Why it Matters How We Frame the Environment» [15] загальним обсягом у 5.142 тис. слововживань. За умовами експерименту зміст статті залишається невідомим для мінімізації дії суб'єктивних факторів на процес моделювання. Вибір тексту наукового стилю пояснюється, насамперед, можливістю спостереження «максимуму семантичного зв'язку» в сполученнях термінологічної лексики, важливої для цього тексту [2, с. 258]. Іншими словами, суттєвими для загального розуміння тексту вважаються найчастотніші терміни або іменники, виділені за допомогою автоматично укладеного частотного словника [4, с. 18]. Релевантними щодо такого тексту вважаються терміни, що зустрілися в тексті не менше двох разів і є семантично зв'язаними з іншими релевантними термінами, тобто утворюють семантичну мережу цього тексту. Для встановлення набору релевантних семантичних компонентів укладено частотний словник словоформ і здійснено лематизацію (див. Табл.1).

Таблиця 1

Найчастотніші терміни статті Лакоффа

№	Термін	Абсолютна частота	Відносна частота
1	frame	72	0,05
2	system	32	0,022
3	brain	25	0,018
4	language	21	0,015
5	framing	21	0,015
6	world	19	0,013
7	word	19	0,013

За результатами аналізу даних Табл. 1 виявляється, що семантична структура тексту організована за ієрархічним принципом, а релевантні компоненти значення є нерівноправними. В ролі текстового концепту – найважливішого для цього тексту поняття – виступає найчастотніший термін *frame*. Зіставлення реєстру найуживаніших термінів з ключовими словами, виділеними автором [15, p. 70] без урахування параметру частоти (*Framing; the Real; Messaging; Enlightenment Reason; Hypocognition; Global Warming*), виявляє збіг лише за однією позицією – *Framing*, і цей термін не є найчастотнішим у нашому тексті. Отже, використання статистичного апарату уможливлює застосування об'єктивних критеріїв до визначення зв'язку між смислом і структурою тексту.

Обмеженість спостереження цілісного тексту в корпусі та незначний обсяг досліджуваного матеріалу зумовлює встановлення зв'язків релевантних семантичних компонентів на підставі даних конкордансів. Для встановлення синтагматичних зв'язків компонентів за розмір вікна конкордансу прийнято повне речення від крапки до крапки. План змісту кожного компонента трактується як множинність синтагматичних відношень із релевантними для цього тексту компонентами. У такий спосіб, у відповідність кожному компоненту ставиться ієрархічно упорядкований за частотою набір релевантних компонентів, зафіксованих в укладених конкордансах. Частота появи компонентів разом характеризує їхню семантичну віддаленість у такомуу тексті й має відповідати встановленому статистичному порогу (≥2). Відповідно до деривaційного критерію семантично зв'язаними вважаємо два слова, якщо одне з них входить у набір семантичних компонентів іншого або слова мають спільний компонент. Аналіз конкордансів кожного релевантного терміна дозволяє встановити його зв'язки в тексті й частоти зв'язаних з ним компонентів. Так, текстовий концепт *frame* входить у п'ять з шести наборів компонентів значення тексту, а складові набору *frame* відрізняються максимальною частотою в конкордансі:

frame(system – 17, brain – 13, word – 9, language – 6),
system (frame – 18, brain – 5, word – 5, world – 3, language – 2, framing – 2),
brain (frame – 13, language – 6, system – 5, word – 2),
world (language – 4, system – 2, word – 2),
framing (system – 3, language – 2, frame – 2),
language (frame – 4, world – 4, system – 3, framing – 2),
word (frame – 15, system – 5, world – 2).

На етапі конкордансингу автоматичний корпусний аналіз тексту завершується. Семантично зв'язаними вважаємо терміни, які входять у набори компонентів один одного, а похідним вважається компонент, який входить у набір іншого компонента. Система семантичних зв'язків тексту презентується у вигляді мовленнєвої, або текстової, мережі – графа, вершини якого представлені компонентами значення тексту, а ребра – семантичним відношенням між ними [2, с. 260–261]. У такому разі ребра текстової мережі позначають синтагматичні відношення між релевантними семантичними компонентами і характеризуються важливою властивістю – рекурсивністю, тобто здатністю до зворотного зв'язку. Нижче подано граф (див. Рис. 1), побудований з урахуванням спаду частот і наборів семантичних компонентів релевантних термінів:

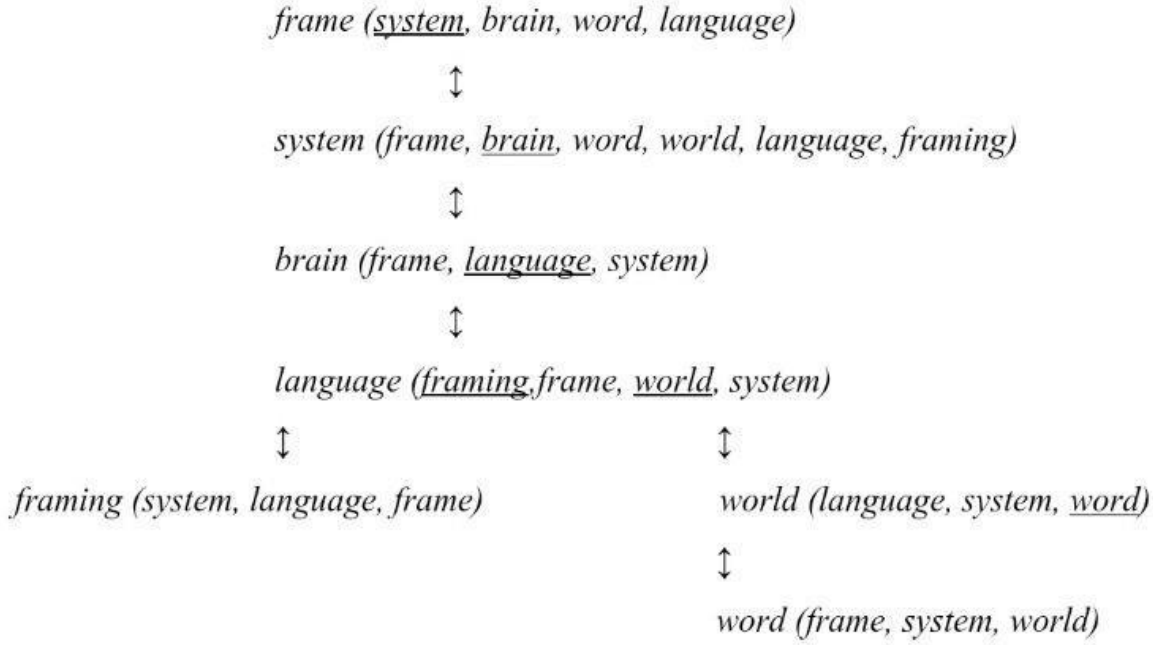

Рис. 1. Семантична модель тексту статті Дж. Лакоффа

Традиційно лексикографічна процедура виділення семантичних складників вважається завершеною, якщо визначення слів або повторюються у словнику, або презентують замкнене коло: такі лексичні значення приймаються за кінцеві семантичні компоненти аналізованого слова [2, с. 245]. На основі аналізу відношень між парами виділених семантичних компонентів встановлюються відношення похідності в напрямку від загального до конкретного. Побудований у напрямку від загального до конкретного поняття і з урахуванням похідності компонентів граф має кільцеву структуру (*world*↔*word*), оскільки набір компонентів найбільш загального поняття *world* включає найбільш конкретне *(language, system, word)*, а набір останнього – *word* містить найбільш загальне поняття *(frame, system, world)*:

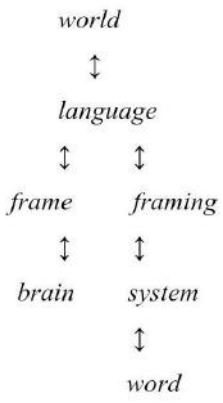

Рис. 2. Семантична модель тексту «від загального до конкретного»

Однак більш оптимальною презентацією структури значення тексту вважаємо просторову мережеву модель. Збільшення обсягу досліджуваного текстового матеріалу (принаймні до 1 млн. слововживань) або використання корпусу текстів уможливить, на нашу думку, виявлення семантичних зв'язків через встановлення колокацій семантичних компонентів тексту. У цьому тексті в ролі такої колокації виступає лише усталене сполучення *system of frame,* що зустрілось 6 разів.

Здійснене дослідження дозволяє дійти таких **висновків**: 1. В основу корпусного підходу до мережевого моделювання тексту покладено концепцію про наявність безпосереднього зв'язку між семантичними та статистичними ознаками слова. 2. Корпусно-базована методика мережевого моделювання передбачає встановлення набору релевантних компонентів значення тексту, поєднаних синтагматичними відношеннями. 3. Послідовність процедур мережевого моделювання тексту включає встановлення набору та ієрархії компонентів значення тексту на підставі текстозорієнтованих словників, їх рекурсивних зв'язків і побудову корпусно-базованої семантичної моделі тексту.

Список використаної літератури

1. Бобкова Т.В. Особливості структури семантичного графа поняття Quantity / Т.В. Бобкова // Науковий вісник Волинського нац. ун-ту ім. Лесі Українки. Філологічні науки. – 2010. – № 8. – С. 143–149.

2. Дарчук Н. П. Комп'ютерна лінгвістика (автоматичне опрацювання тексту) / Наталія Петрівна Дарчук. – К. : ВПЦ «Київський університет», 2008. – 351 с.

3. Дарчук Н.П. Напрямки формалізації семантики / Н.П. Дарчук // Мовні і концептуальні картини світу. – 2013. – Вип. 46, Ч. 1. – С. 385–396.

4. Дарчук Н.П. Лінгвістичне забезпечення автоматичних систем аналізу українськомовного тексту (на прикладі системи автоматичного граматичного аналізу тексту АГАТ) : автореф. дис. … доктора філол. наук : 10.02.01, 10.02.21/ Н.П. Дарчук; Київ. нац. ун-т ім. Т. Шевченка. – К. : [б. в.], 2015. – 34 с.

5. Касевич В.Б. Буддизм. Картина мира. Язык // Вадим Борисович Касевич. – СПб : Центр «Петербургское востоковедение», 1996. – 278 с.

6. Перебийніс В.І., Бобкова Т.В. Частота мовних одиниць як відображення їхніх системних характеристик / В.І. Перебийніс, Т.В. Бобкова // Проблеми загального, германського та слов'янського мовознавства. – Чернівці : Книги-ХХІ, 2008. – С. 446–453.

7. Скороходько Э.Ф. Семантические сети и автоматическая обработка текста / Эдуард Федорович Скороходько. – К. : Наук. думка, 1983. – 217 с.

8. Тулдава Ю. Проблемы и методы квантитативно-системного исследования лексики : [монография] / Юхан Тулдава. – Таллин : Валгус, 1987. – 204 с.

9. Baroni M., Murphy B., Barbu Ed., Poesio M. Strudel : A Corpus-Based Semantic Model Based on Properties and Types / M. Baroni, B. Murphy, Ed. Barbu, M. Poesio // Cognitive Science. – 2010. – № 34. – P. 222–254.

10. Fourtassi A., Dupoux E. A Corpus-based Evaluation Method for Distributional Semantic Models / A. Fourtassi, E. Dupoux // Proceedings of the ACL Student Research Workshop, Sofia, August 4-9, 2013. – Sofia : Association for Computational Linguistics, 2013. – P. 165–171.

11. Murphy B., Talukdar P., Mitchell T. Selecting Corpus-Semantic Models for Neurolinguistic Decoding / B. Murphy, P. Talukdar, T. Mitchell // Proceedings of First Joint Conference on Lexical and Computational Semantics. – Montreal : 2012. – P. 114.

12. Mihalcea R., Corley C., Strapparava C. Corpus-based and Knowledge-based Measures of Text Semantic Similarity // R. Mihalcea, C. Corley, C. Strapparava // American Association for Artificial Intelligence. – 2006. – P. 775–780.

13. Tefera A., Assabie Y. Automatic Construction of Amharic Semantic Networks From Unstructured Text Using Amharic WordNet / A. Tefera, Y. Assabie // Proceedings of the Seventh Global Wordnet Conference. – Tartu, 2014. – P. 172–177.

14. Tognini-Bonelli E. Corpus Classroom Currency / E. Tognini-Bonelli // Darbai ir Dienos. – 2000. – No 24. – P. 205–243.

Джерела ілюстративного матеріалу

15. Lakoff G. Why it Matters How We Frame the Environment / G. Lakoff // Environmental Communication. – 2010. – Vol. 4. – No. 1. – P. 70–81.

Аннотация

Бобкова Т. В. Корпусно-ориентированная методика сетевого моделирования текста.

В работе анализируются подходы к сетевому моделированию значения текста, основанные на использовании классификаций предметов внешнего мира и различающиеся методами и приемами моделирования. Обосновывается необходимость и перспектива использования корпусно-ориентированного подхода, способствующего автоматизации анализа семантики текста на основе статистического и корпусного метода.

Установлены основные этапы сетевого моделирования семантики текста на базе корпусно-ориентированной методики. Предлагаемая методика сетевого моделирования, основанная на применении статистического анализа и извлеченных из текста конкордансов, способствует упорядочению объективных данных о семантических связях компонентов значения текста. Сочетание статистического и корпусного методов способствуется построению пространственной трехмерной модели значения текста.

***Ключевые слова**: сетевое моделирование, текст, значение, семантический компонент, корпусно-ориентированный, частота.*

Summary

Bobkova T. V. Corpus-based methodology of text network modelling.

The approaches to the semantic network modelling of the text, which are based on the outer world subjects classification and can be distincted by modelling methods and techniques, are analysed in the article. The necessity and prospects of using corpus-based approach, which makes possible the automatization of semantic text analysis with help of statistic and corpus methods, is proved.

The main stages of network modelling of text semantics according to the corpus oriented methodology are determined. The suggested network modelling methodology, which is based on applying statistic analysis and the extracted from the text concordances, helps managing the objectivized data about the semantic relations between the text meaning components. The combination of statistic and corpus methods furthers the construction of the spacial three-dimensional model of text meaning.

***Keywords**: semantic network modelling, text, meaning, semantic component, corpus-based, frequency.*