Створення лексико-тематичної моделі дискурсу на основі синтаксичних критеріїв

Наталія Чейлитко[1], Андрій Галкін[2]

1) Лабораторія комп'ютерної лінгвістики, Київський національний університет імені Тараса Шевченка, УКРАЇНА 02154, м. Київ, вул. Русанівська набережна, 10, кв. 37. E-mail: natalia.cheilytko@gmail.com

2) Відділ архітектури і розробки програмних продуктів, ТОВ Арт-Мастер, УКРАЇНА 02140, м. Київ, вул. Вишняківська 6а, кв. 298, E-mail: andy.galkin@gmail.com

Abstract* – *The article is devoted to the correlation of semantic and syntactic text structures. A new approach for compousing the lexico-thematic model of discourse is represented.

Keywords – lexico-thematic model of discourse, lexico-thematic tree, meaning, sentence syntactic structure, zone of word form's connections.

Виявлення смислу в текстах, які репрезентують певний дискурс, є однією з наукових проблем, над якою лінгвісти працюють вже протягом кількох десятиліть. Незважаючи на те, що науковцями запропоновано чимало підходів до її розв'язання, проблему, напевно, можна вважати вічною. Адже, як влучно зауважив В. А. Звегінцев, семантика є тією туманною сферою, «в яку занурювалося чимало сміливих дослідників – деякі з них так і загубилися в ній, а ті, хто повертався, зазвичай не приносили з собою значних результатів» [1, с. 80].

Завдяки розвитку комп'ютерних технологій можна верифікувати теоретичні лінгвістичні моделі шляхом створення на їх основі систем автоматичного опрацювання тексту (АОТ) та застосування цих систем для аналізу текстів певної мови. Настанова на практичне використання певної теоретичної моделі стимулює розвиток лінгвістики, зокрема відбувається уточнення відомих лінгвістичних теорій і формулювання нових, усвідомлення потреби в нових лінгвістичних понятях, вироблення нових теоретичних концепцій. Таким чином, створюються передумови нової наукової дисципліни – експериментальної лінгвістики [2].

Семантичні моделі мають різну складність залежно від глибини відображення семантичних процесів у дискурсі. Моделі, які відтворюють тематичне розмаїття дискурсу, узагальнено відображають його смисл. Такі моделі покликані дати відповідь на питання «про що, коли, кому, ким і з якою метою повідомляється».

Актуальним у цьому випадку постає завдання виробити послідовну методику створення тематичної моделі дискурсу. Продуктивним шляхом створення тематичної моделі дискурсу є підхід, який базується на виявленні в тексті тематично значущих ЛСВ та групування їх в лексико-тематичні групи, що стають підґрунтям моделювання смислу дискурсу на рівні його теми, оскільки, як слушно зазначив В. С. Баєвський, кожне повнозначне слово у тексті становить тему і тому близькі за значенням слова в межах певного тексту утворюють великі сюжетні теми [3].

У праці, присвяченій формалізованому виявленню семантичної відстані між словниками двох текстів, А. І. Новіков та О. І. Ярославцева припустили, що значущість певного ЛСВ для смислового розгортання повідомлення визначається кількістю його семантичних зв'язків з іншими ЛСВ [4]. Тієї ж думки дотримується інший дослідник – Е. Ф. Скороходько, який розробив механізми автоматичного опрацювання тексту на основі семантичних мереж [5].

Прикладне спрямування названих двох досліджень стимулювало їхніх авторів до пошуку формальних критеріїв встановлення семантичних зв'язків між словоформами. Підґрунтям став закон семантичного узгодження, сформульований В. В. Гаком: «будь-які два слова, які пов'язані синтаксично, завжди пов'язані й семантично (зворотне в загальному випадку неправильно)» [5, с. 103]. Згідно з таким поглядом, речення розглядається як формальний ланцюжок, що перебуває в певній залежності від смислового ланцюжка. Тому в ролі формального критерію семантичного зв'язку між словами в тексті визнано синтаксичний зв'язок. Тоді виявляється, що для встановлення семантичного зв'язку між словоформами в тексті виникає потреба в попередньому синтаксичному аналізі, який покликаний відобразити кількість синтаксичних зв'язків кожного ЛСВ.

Методом, який враховує синтаксичні й позиційні властивості речення й тексту, є метод зон зв'язків словоформ. **Зоною зв'язків словоформи (ЗЗС)** називають ту частину речення, у якій реалізуються всі синтаксичні зв'язки словоформи. ЗЗС визначаються з урахуванням особливостей входження словоформи до складу певного члена речення.

Застосування методу ЗЗС полягає в розгляді синтаксичної структури речення в таких аспектах.

1. Синтагматичний аспект передбачає аналіз реалізованого в мовленні речення як лінійної послідовності його мінімальних компонентів – елементів речення. У письмовому варіанті тексту **елемент речення (ЕР)** дорівнює сукупності буквених символів між двома пробілами.

2. Структурний аспект полягає в тому, що речення представляється у вигляді сукупності синтаксичних зв'язків між його компонентами. Структуру речення розглянуто як організовану за допомогою синтаксичних зв'язків ієрархію ЕР, представлену у вигляді **дерева залежностей**. Вузли дерева відповідають елементам речення, ребра – зв'язкам між ЕР.

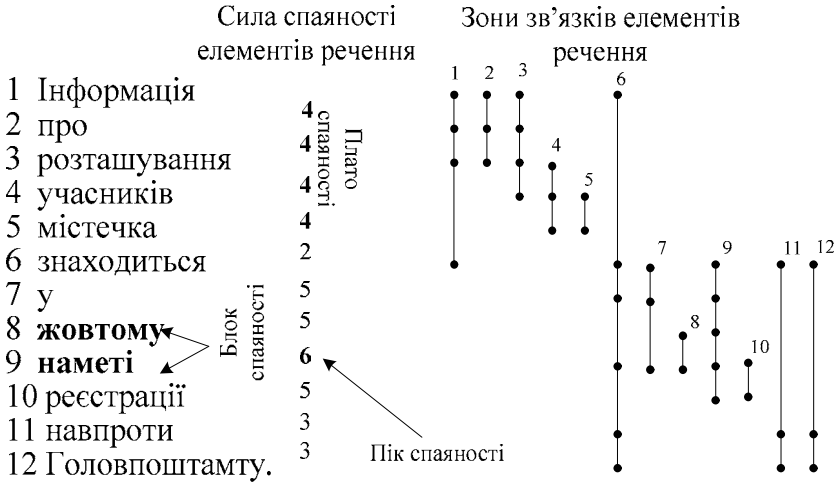

Рис. 1. Піки, плато та блоки спаяності елементів речення

ЗЗС відображаються графічно шляхом проведення відрізка через усі словоформи, пов'язані з аналізованою. Представивши в такий спосіб зони зв'язків для всіх елементів речення, стає можливим обчислити, скільки ЗЗС припадає на інтервал між кожними двома елементами речення, лінійно розташованими поряд, тобто визначити силу їхньої синтаксичної спаяності. На Рис. 1 відображені ЗЗС для всіх дванадцяти елементів речення: в інтервалах між кожними двома елементами, розташованими поряд, наведено числовий показник, що дорівнює кількості ЗЗС, які містяться в цьому інтервалі (на рисунку кожну зону зв'язків позначено числом, яке вказує на порядковий номер у реченні того ЕР, для якого визначена ця зона). Цей показник називають **силою спаяності ЕР**. Чим більша кількість ЗЗС припадає на інтервал між двома елементами речення, тим більш зв'язними в лінійному потоці мовлення вважаються ці елементи. Найбільший показник сили спаяності в реченні називають **піком спаяності ЕР**. Якщо кілька послідовно розташованих показників мають однакове найбільше значення, таку послідовність називають **плато спаяності ЕР**. У реченні може бути кілька піків та плато спаяності. Елементи речення, між якими сила спаяності максимальна, утворюють **блок спаяності ЕР**.

Зони зв'язків будуються для всіх елементів речення, незважаючи на те, є вони синтетичною словоформою чи частиною аналітичної словоформи, репрезентують повнозначні чи службові словоформи, бо нас цікавила сила «зчеплення» між кожними двома лінійними сегментами речення, розділеними пробілом.

Блоки спаяності є осередками максимальної синтаксичної зв'язаності елементів речення, отже, їх можна розглядати як сильні позиції речення. Разом з тим, на провідній ролі блоків спаяності не лише в позиційній, але й семантичній організації речення наголошується в колективній монографії «Закономірності структурної організації науково-реферативного тексту» [6]. Автори припустили, що до блоків найвищої спаяності ЕР потрапляють ті ЛСВ, які мають найбільше смислове навантаження в межах речення.

Ми вирішили перевірити цю гіпотезу. Узявши за основу метод ЗЗС, довели, що сукупність усіх ЛСВ, які ввійшли до блоків спаяності у текстах аналізованого дискурсу, адекватно відобразили тематичну спрямованість цього дискурсу.

Матеріалом для проведення експерименту стала сукупність текстів публікацій газети «Українська правда», присвячених виборам 2004 року (3217 речень).

З метою оптимізації та уніфікації процесу визначення ЗЗС розроблено спеціальне програмне забезпечення, завдяки якому створено електронну базу дерев залежностей, а також автоматично побудовано та проаналізовано зони зв'язків усіх словоформ на основі дерев залежностей (Рис. 2).

Зазначимо, що набір правил для обчислення ЗЗС може формуватися динамічно, відповідно до характеру дослідження. У цьому дослідженні розроблено 2 типи правил.

1. Правила, які мають спрацьовувати у тих випадках, якщо знайдено вузол дерева, що відповідає певному типу ЕР.

2. Правила, які мають спрацьовувати за умови, що знайдено ребро, марковане певним способом.

Здійснюючи рекурсивний перебір усіх вузлів дерева, починаючи з вершини дерева, алгоритм перевіряє, чи відповідає частина дерева певній умові-ситуації. Якщо знайдено таку відповідність, то алгоритм виконує передбачену конкретною ситуацією сукупність дій, які полягають у додаванні певного вузла до зони іншого вузла.

Основні дії алгоритму: 1) знайти зону зв'язків, побудовану для певного вузла; 2)додати вузол у зону іншого вузла дерева.

Зауважимо, що алгоритм належить до так званих rule-based system [7], тому сукупність правил, які лежать в основі алгоритму, не становить ієрархії та не передбачає якоїсь чіткої, наперед визначеної послідовності виконання.

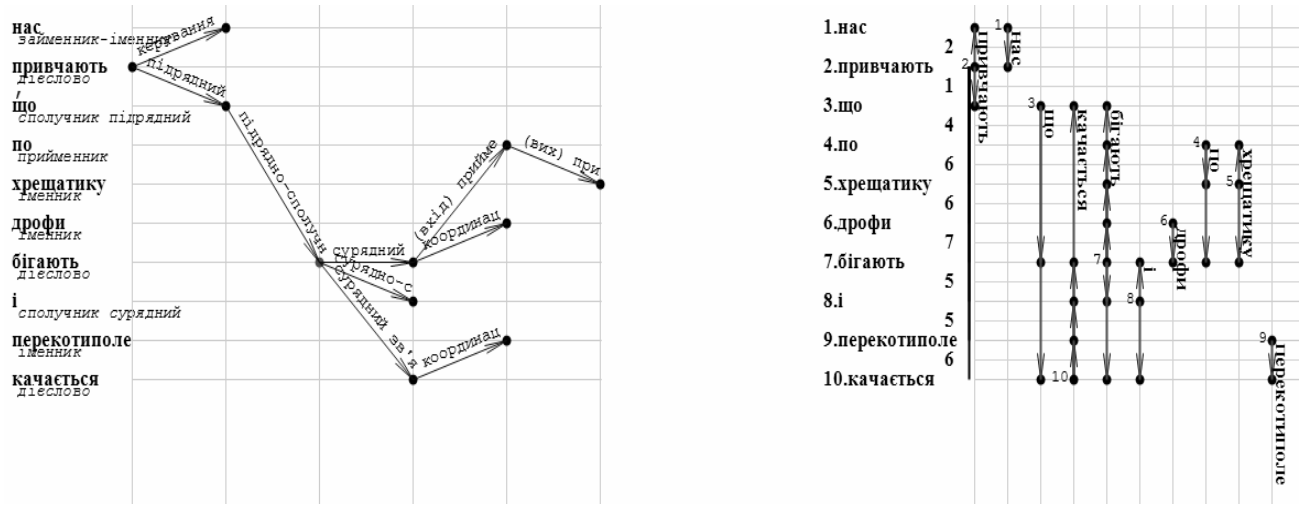

Рис. 2. Приклад автоматичного визначення зон зв'язків словоформ на основі дерева залежностей

Таким чином, метод ЗЗС став підґрунтям для укладання лексико-тематичної моделі дискурсу політичних новин, присвячених виборам 2004 року.

Свого часу Н. Ю. Шведова показала, що «картинки життя» в ідеографічному дереві «вимальовуються», насамперед, лексикою, що називає різноманітні реалії буття, – іменниками та дієсловами [8]. Тому ми уклали два лексико-тематичні дерева тих менників та дієслів, які ввійшли до складу всіх блоків спаяності в проаналізованих текстах. Ми припустили, що лексико-тематична модель (ЛТМ), яка становить сукупність лексико-тематичних дерев (ЛТД) повнозначних частин мови, адекватно відобразить «картину життя», що постає з текстів політичних новин, присвячених виборам 2004 р.

Процес створення лексико-тематичної моделі складався з таких основних етапів:

1) відбір елементів речення, які входять до всіх блоків спаяності, виявлених у досліджуваній вибірці текстів;

2) лематизація (приведення до словникових форм) та впорядкування ЕР за частиномовною віднесеністю;

3) визначення для кожної леми (словникової форми) її значення на основі вивчення контекстів її вживання – відповідно, надалі у фокусі уваги перебувають не леми, а лексико-семантичні варіанти (ЛСВ);

4) присвоєння кожному ЛСВ числового індексу, який відображає кількість випадків, коли аналізований ЛСВ увійшов до складу блоку спаяності;

5) впорядкування ЛСВ, які належать до одного лексико-граматичного класу слів, у лексико-тематичне дерево (ЛТД).

Блоки спаяності визначено за правилом: якщо певний показник сили спаяності перевищував два інші, сусідні з ним показники, або дорівнював їм, то такий показник вважали піком спаяності. Елементи речення, між якими виявлено пік спаяності, утворюють блок спаяності.

У ході групування ЛСВ іменників у лексико-тематичне дерево ми спиралися на структуру «Російського семантичного словника» за редакцією Н. Ю. Шведової [9], а під час створення відповідного дерева дієслів – на «Тлумачний словник дієслів» за редакцією Л. Г. Бабенко [10]. Значення лексем встановлювалися із залученням тлумачного словника української мови [11].

Структуру лексико-тематичного дерева описано через рівні глибини дерева: ЛТД розглядається як ієрархія вузлів (назв тематичних полів/груп тощо) різних рівнів. Відповідно, ЛТД становить ієрархію, організовану за принципом: від цілого до його частини.

У ЛТД міститься кількісна характеристика кожного ЛСВ та кожного вузла – індекс (**I**). Індекс показує, скільки разів ЛСВ увійшло до складу всіх блоків спаяності, визначених у проаналізованих текстах.

На жаль, через невеликий обсяг цієї публікації ми не зможемо детально описати створену лексико-тематичну модель дискурсу політичних новин, присвячених виборам 2004 р., тому обмежимося лише загальною характеристикою побудованих лексико-тематичних дерев іменників і дієслів та подамо найяскравіші, на наш погляд, ілюстрації.

Загальна кількість ЛСВ, які потрапили до ЛТД іменників, становить 1740 одиниць. Глибина дерева сягає шести рівнів. Від вершини дерева відходить п'ять ребер. Відповідно, п'ятьома основними вузлами дерева є вузли другого рівня, які найбільш загально окреслюють явища дійсності, названі іменниками в проаналізованому дискурсі: «Буття», «Абстрактні відношення», «Людина», «Суспільство», «Природа». Найбільш наповненими лексико-семантичними варіантами є вузли «Людина» – 42,53 % від загальної кількості ЛСВ, «Суспільство» – 41,32 % ЛСВ. Ці ж два вузли характеризуються найбільшими показниками індексу, тобто частотою потрапляння ЛСВ до блоків спаяності, – 41,56 % для «Людини» і 42,99 % для «Суспільства». Цей факт яскраво свідчить про основну тематичну спрямованість дискурсу – суспільне життя людини.

Л. Ставицька писала, що під час подій, які розгорталися навколо виборів 2004 року, відбувалося стихійне об'єднання людей у спільноти, коли «в очах один одного прочитується та сама думка, а спілкування робить «спільність людей» не фантомною»

[12]. Як зауважує дослідниця, тоді ж відбулася помітна активізація слова *нація*, що витіснило на периферію слово *народ,* «висока експресія якого на сьогодні помітно здевальвована жонглюванням ним з боку політиків різних мастей» [12]. Матеріал газети «Українська правда» засвідчує протилежне: ЛСВ *народ* (**I**=78) має значно більший індекс за ЛСВ *нація* (**I**=8). Це наводить на думку про завчасність визнання за словом *народ* периферійного статусу в тогочасному політичному дискурсі, принаймні у висвітленні названої газети. Хоча теза про зниження його експресивної сили видається цілком слушною.

Піддерево «Людина» містить чотири розгалуження: «Людина як соціальна істота», «Людина як фізична істота», «Людина як психічна істота», «Людина як розумна істота». Значно переважають показники сумарного індексу по піддереву першого вузла (10,19 %), що вказує на тематичну домінанту дискурсу – соціальну діяльність людини.

У піддереві вузла «Людина як психічна істота» наявні ЛСВ, які дають уявлення про психологічне тло подій, відображених у текстах, – настрої та світовідчуття людей. Так, невипадковою є назва статті Л. Ставицької, присвяченої аналізу тогочасного політичного дискурсу, – «Дискурс помаранчевої пристрасти» [12]. Таку рису, як пристрасність виборів 2004 відображено й в ЛТД іменників: у вузлі «Емоційний стан» переважають ЛСВ на позначення бурхливих людських переживань (*вогонь ‘душевне піднесення, натхнення’* (**I**=1); *ейфорія ‘сприйняття, оцінка чого-небудь у надто або невиправдано оптимістичних тонах’* (**I**=1); *захоплення ‘велике внутрішнє піднесення, збудження, порив, запал’* (**I**=1); *істерика ‘напад істерії’* (**I**=2); *обурення ‘сильне невдоволення, роздратування’* (**I**=1); *підйом ‘ентузіазм’* (**I**=1); *піднесення ‘запал, натхнення, душевне піднесення, ентузіазм’* (**I**=1); *пристрасть ‘сильне, бурхливе, нестримне у своєму виявленні почуття’* (**I**=2); *раж ‘сильне збудження, несамовитість, шаленство’* (**I**=1); *хвиля ‘наплив яких-небудь почуттів, думок, що визначають настрій людини’* (**I**=1)).

До лексико-тематичного дерева дієслів увійшло 902 ЛСВ. Максимальна глибина дерева – п'ять рівнів. Дерево має три розгалуження від вершини – вузли «Дія та діяльність», «Буття, стан, якість», «Відношення». Значна кількість ЛСВ з високими показниками індексу в першому з трьох піддерев – «Дія та діяльність» (сумарний індекс по піддереву – 58,06 %) засвідчує те, що проаналізовані тексти, насамперед, присвячені відображенню активних дій, а не станів або процесів.

Тому, на нашу думку, цей дискурс правомірно віднести до діяльнісно-прагматичного типу політичних дискурсів. Вивчивши, які ЛСВ найбільше представлені в ЛТД дієслів, можна стверджувати, що в текстах, які репрезентують досліджуваний дискурс, персонажі мислять, спілкуються, пересуваються й досягають мети.

Таким чином, нами підтверджено гіпотезу про можливість створення лексико-тематичної моделі дискурсу на основі блоків спаяності ЕР. Упорядковані в лексико-тематичну модель ЛСВ іменників та дієслів адекватно відобразили тематичну спрямованість дискурсу політичних новин, присвячених виборам 2004 року.

Убачаємо великий потенціал в описаній методиці, оскільки на її основі стає можливим проаналізувати різноманітні дискурси, визначити домінантні та периферійні теми, виявити особливості оцінки авторами текстів відображуваних подій.

References

[1] В. А. Звегинцев, Предложение и его отношение к языку и речи. М. : Эдиториал УРСС, 2001, 312 с.

[2] Лингвистическое обеспечение системы «Этап–2» / [Ю. Д. Апресян, И. М. Богуславский, Л. Л. Иомдин и др.], М. : Наука, 1989, 296 с.

[3] В. С. Баевский, Лингвистические, математические, семиотические и компьютерные модели в истории и теории литературы, М. : Языки славянской культуры, 2001, 336 с.

[4] А. И. Новиков, Е. И. Ярославцева, Семантические расстояния в языке и тексте, М. : Наука, 1990, 136 с.

[5] Э. Ф. Скороходько, Семантические сети и автоматическая обработка текста, К. : Наук. думка, 1983, 212 с.

[6] Закономерности структурной организации научно-реферативного текста / [Л. М. Гриднева, Т. А. Грязнухина, Н. П. Дарчук и др. ; отв. ред. В. И. Перебейнос], К. : Наукова думка, 1982, 322 с.

[7] Rule-Based Systems and Identification Trees [Електронний ресурс], Режим доступу до статті : http://ai-depot.com/Tutorial/RuleBased.html.

[8] Н. Ю. Шведова, Русский язык: Избранные работы, М. : Языки славянской культуры, 2005, 640 с.

[9] Русский семантический словарь. Толковый словарь, систематизированный по классам слов и значений / [под общей ред. Н. Ю. Шведовой], М. : Азбуковник, 1998-2003, Т. 1–3.

[10] Толковый словарь русских глаголов : Идеографическое описание. Английские эквиваленты. Синонимы. Антонимы / [под ред. проф. Л. Г. Бабенко], М. : АСТ-ПРЕСС, 1999, 704 с.

[11] СЛОВНИК.НЕТ. [Електронний ресурс], Режим доступу до словника : http://slovnyk.net/.

[12] Л. Ставицька, Дискурс помаранчевої пристрасти [Електронний ресурс], Режим доступу до статті : http://www.textology.ru/public/pomaranch.html.