УДК 10.02.21

С.В. Морін,
викладач
(Житомирський державний технологічний університет)

КОЛИВАННЯ МОВЛЕННЄВОЇ ЕНТРОПІЇ ПРИ АНАЛІЗІ СТРУКТУРИ АНГЛІЙСЬКИХ І УКРАЇНСЬКИХ ПУБЛІЦИСТИЧНИХ ТЕКСТІВ У ТЕРМІНАХ ДИСТРИБУЦІЇ

У статті досліджено структура тексту шляхом аналізу коливань мовленнєвої ентропії на прикладі англійських та українських публіцистичних текстів за допомогою дистрибутивної синтактико-семантичної моделі. Здійснено порівняльний аналіз структур англійських та українських текстів. Методом дистрибуції доказано поділ текстової структури на мікроблоки, а також наявність змістовно насичених речень у структурі тексту.

Текст ─ це структурна єдність, що виявляється у взаємозв'язку і взаємодії змістовної і лексичної структур. Такій єдності притаманні багато рівнів (звуковий, лексичний, морфологічний, прийменниковий), окрім лінійності (тема / рема, анафоричність / катафоричність, лексичні повтори, перша / не перша фраза, вираження визначеності / невизначеності й под. [1].

Завдяки багатоплановості текст ─ це складний об'єкт дослідження. Цікавим є дослідження такого об'єкта за допомогою формальних методів. З відомих нам формально-семантичних методів аналізу (дескрипторний, синсемний, дистрибутивний та ін.) [2], ми звернулися до дистрибутивного.

Цей метод заснований на математичній теорії множин. Текст розглядається як структура, що складається з елементів семантико-синтаксичних множин у контексті *дистрибуції*, суть якої можна коротко охарактеризувати як *закономірний розподіл і сполучення* елементів (частин слів, слів і словосполучень) у тексті [3; 4], виступаючи при цьому непрямим експлікантом значення [5]. Більш докладно такий підхід розроблений і представлений у докторський дисертації Г.Е. Мирама "Дистрибутивная модель синтаксиса и семантики научно-технического текста для систем автоматической обработки информации" у вигляді дистрибутивної синтактико-семантичної моделі (ДССК) [4].

Таким чином, лінійність текстової структури може бути представлена у вигляді послідовності дистрибутивно-семантичних класів, здатних виявляти змістовну (інформативну) сторону текстової структури.

У роботі аналізується величина мовної ентропії, яку розуміємо як ступінь невизначеності лексичної сполучуваності слова, вираженої через апарат моделі ДССК.

Ціль роботи – *аналіз категорії регулярності / нерегулярності коливань мовленнєвої ентропії як міри невизначеності лексичної сполучуваності з метою вивчення потенційних можливостей моделі ДССК відображати змістовну (інформаційну) структуру тексту.*

В основі моделі ДССК лежать наступні ключові параметри:

- семантичні об'єкти (СМО, СМО*),
- синтаксичні об'єкти (СНО, СНО*),
- семантичні відносини.

Об'єкти й відносини встановлюються за допомогою діагностуючих мовленнєвих оточень, що утворюють *дистрибутивно-семантичні класи* (ДсмК) слів [4].

Спочатку розглянемо семантичний рівень моделі. *Семантичним об'єктом* моделі (СМО) є елемент ДсмК *іменників*. Належність до одного або декількох ДсмК ділить СМО іменників на однозначні та багатозначні. При цьому СМО може містити в собі один або цілий ряд *лексико-семантичних варіантів* (ЛВС), які розглядаються як один СМО, наприклад, *collection (P)*– нагромадження, зібрання, збори; *work (P)* – робота, праця, заняття; *canvas (O)* – полотнище, парусина, канва.

Діагностуючі мовленнєві оточення для визначення ДсмК англомовних текстів представлені у роботі Г.Е. Мірама [4]. Для української мови ми розробили власні діагностуючі оточення, доповнивши їх новими класами ДсмК ─ "конкретні, незлічувальні іменники", "людина", "закритий простір", "відкритий простір". ДсмК іменників англійської та української мов позначаються індексами – *O (речи), P (процеси), C (характеристики), S (речовина), M (синтаксеми), H (людина), R (закритий простір), T (відкритий простір), які визначаються за допомогою* ***первинних*** *діагностуючих оточень методом підстановки.* Для української мови передбачений додатковий ДсмК "конкретні, незлічувані іменники" (індекс *B*).

Другою групою семантичних об'єктів (СМО) є група ДсмК* відносних і абсолютних атрибутів, дієслів (перехідні та неперехідні), дієприкметників, прислівників, прийменників, *які визначаються за допомогою* ***вторинних*** *мовленнєвих оточень методом підстановки.*

Таким чином, у термінах моделі ДССК план змісту англійського й українського публіцистичного тексту представлений двома типами семантичних об'єктів, СМО й СМО*, де *СМО моделюють номінативний компонент, а СМО* ─ його атрибутивний компонент*.

Наприклад, наступне речення "Недавно у Великобританії пройшов пілотний аукціон з торгівлі квотами на викид шкідливих парникових газів" у термінах СМО й СМО* може бути представлене:

/*/ /*/ (T) /PMH/ /OPM/ (P) P/*/ (P) (C) /*/ (CSP) /*/ /ASTR/ (S).

де /*/ ─ СМО* атрибутів універсальної семантики, /PMT/ ─ СМО* атрибутів конкретної семантики, (P) ─ СМО іменників конкретної семантики.

На прикладі англійської мови речення "Ukraine was not among the bidders, thus letting billions of dollars slip through its fingers" має такий вигляд:

(T) /*/ &(not) /*/ &(the) (H) /*/ /*/ &(billions) /*/ (O) /*/ /*/ (O).

де /*/ ─ СМО* атрибутів універсальної семантики, /O/ ─ СМО* атрибутів конкретної семантики, (Р) ─ СМО іменників конкретної семантики, &(the) ─ семантичні об'єкти, які не реєструються моделлю ДССК.

СМО іменників можуть вступати у відповідні *семантичні відносини* з іншими іменниками в *субстантивних безприйменникових словосполученнях*. Причиною дослідження параметра семантичних відносин є необхідність аналізу семантики в таких "ланцюжках" СМО, що складаються із двох, трьох і більше іменників, що характерно для англійської мови.

Таким чином, наступним параметром моделі ДССК запроваджується п'ять семантичних відносин ─ "визнавати за суб'єкт", "визнавати за об'єкт", "визнавати за місце", "визнавати за інструмент", "визнавати за міру", запозичених з роботи Г.Е. Мірама [4], для визначення котрих використаються діагностуючі оточення із двома змінними елементами. Це ─ тестуюче слово (іменник) і типовий представник ДсмК. Семантичні відносини встановлюються між тестуючим словом і типовим представником ДсмК.

У процесі визначення відносини "визнавати за суб'єкт", "визнавати за об'єкт" підключається етап трансформаційного аналізу (ТА). Ці відносини встановлюються тільки для предикатів, і завдання ТА – замінити нечітко виражені семантичні відносини іменника більш конкретними дієслівними валентностями. При встановленні відносини "визнавати за суб'єкт" іменник-предикат перетвориться у відповідне дієслово дійсного стану. При встановленні відносини об'єкта іменник трансформується у відповідне дієслово пасивного стану.

Як приклад визначимо семантичні відносини, які можуть бути встановлені між словом *translation* – переклад (ДсмК "процеси", "речі").

1. Відношення "визнавати за суб'єкт" *translation* може встановлювати з елементами ДсмК "людина" і "пристрій": *a human being (apparatus) translates ─ людина (апарат) перекладає.*

2. Відношення "визнавати за об'єкт" може бути встановлено з елементами ДсмК "речі", "пристрій": *word is (not) translated ─ (не) перекладають слово; machine translation ─ переклад виконується за допомогою машини.*

3. Відношення "визнавати за місце" встановлюється з усіма елементами ДмсК: *object, apparatus, substance, process, value, idea, money, worker, country, room occur in translation ─ речи, пристрій, речовина, процес, величина, ідея, гроші, робітник, країна, кімната зустрічаються в перекладі.*

4. Відношення "визнавати за міру" встановлюється з елементами ДсмК "процеси", "характеристики", "ідея": *process, value, idea are used as a measure of translation ─ процес, характеристика, ідея виступають мірою перекладу.*

5. Відношення "визнавати за інструмент" може бути встановлене з елементами ДсмК "речи", "пристрій", "людина": translation is used with the help of word, apparatus, human being ─ переклад виконується за допомогою слова, пристрою, людини.

Синтаксичний рівень моделі ДССК представлений наступними синтаксичними об'єктами дистрибутивно-синтаксичних класів (СНО): іменники / особисті займенники *(індекс N)*; відносні атрибути *(індекс RA)*; абсолютні атрибути *(індекс AA)*; перехідні дієслова *(індекс TV)*; неперехідні дієслова *(індекс VI)*; прийменники *(індекс P)*; прислівники *(індекс F)*.

Синтаксично незначущі елементи тексту (СНО*) містять у собі ряд елементів, які не описуються моделлю ДССК. До таких елементів належать артиклі, службові дієслова, текстові оператори (напр., which, there, який, що), невизначені займенники (напр., somebody, somewhat, хто-небудь, який), частки (напр., to, бодай, же), модальні дієслова (напр., may, can, могти), квантифікаторы (напр., twice, very, only). Основною відмінністю СНО й СНО* є "десемантизованість" СНО* у рамках дистрибутивної моделі.

Наприклад, речення "Недавно у Великобританії пройшов пілотний аукціон з торгівлі квотами на викид шкідливих парникових газів" у термінах СНО й СНО* може бути представлене подібним чином:

F P N VI RA N P N N P N RA RA N.

де N, P, F, VI, RA ─ дистрибутивно-синтаксичні класи.

На прикладі англійської мови речення "Only Ukraine was not among the bidders, thus letting billions of dollars slip through its fingers" має такий вигляд у термінах СНО й СНО*:

&(only) N VI P N, F VT P N VI P N.

де N, VI, P, F ─ дистрибутивно-синтаксичні класи, &(only) ─ синтаксично незначущі елементи тексту.

Для проведення дослідження були взяті шість ідентичних за змістом текстів (оригінал та переклад на українську) із журналу The Ukrainian, 3/2002: "Ars Longa, Vita Brevis", "$ For Fresh Air", "2001 Regions' Investment "Podium", "Життя коротке, мистецтво ─ вічне!", "Отримай "зелені" за свіже повітря", "Інвестиційний "п'єдестал пошани" регіонів-2001".

Аналіз текстів за допомогою моделі ДССК показав, що більша частина речень починається з невизначених ДсмК (60-77 %) і закінчується визначеними (79-92 %), тобто на семантично невизначених елементах ентропія досягає певного кінцевого рівня й вимагає наступного розгортання акту мовленнєвої комунікації. У той час як на семантично визначених елементах ентропія, навпаки, досягає практично нескінченного рівня й не вимагає продовження або може мати нескінченне число продовжень [4]. Подібне коливання ентропії на початку й кінці англійського й українського речень є регулярним і береться нами за основу як *принцип регулярного коливання*.

Таким чином, дистрибутивна семантико-синтаксична структура англійського й українського текстів складається із чотирьох типів речень:

- починаються з невизначених елементів і закінчуються визначеними;
- починаються з визначених елементів і закінчуються невизначеними;

- починаються й закінчуються визначеними елементами;
- починаються й закінчуються невизначеними елементами.

Якщо принцип регулярного коливання визначених і невизначених елементів початку й кінця речень є правилом, то винятки із правила являють собою наступне: англійський та український тексти мають у складі своїх макроструктур речення, які або *починаються з визначених елементів ДсмК, або закінчуються невизначеними елементами ДсмК.* Для аналізу представляється цікавим місце "стиків" елементів ДсмК подібних речень із елементами попередніх або наступних речень.

Виділивши початкові й кінцеві з'єднання речень за допомогою елементів моделі, відзначимо, що структура тексту ділиться на блоки, які закінчуються групою речень із характерними "стиками" невизначеного елемента ДсмК кінця попереднього речення та невизначеного елемента ДсмК початку наступного. При цьому присутні чітке "затухання" мовленнєвої ентропії на "стиках" та порушення принципу регулярного коливання визначених / невизначених елементів початку й кінця речень.

Подібних "стиків", що складаються як мінімум із двох речень, може бути кілька. Останнє речення групи *завжди* закінчується визначеним елементом ДсмК, вимагаючи або продовження мовленнєвого акту (сигналізуючи про зміну мікроструктури й перехід до нової), або не вимагає продовження (указуючи на закінченість ідеї). Такі групи речень виступають ознаками тематичної невизначеності, за яких відбувається, на наш погляд, зміна мікротеми тексту. Подібний розподіл структури тексту є регулярним і виступає непрямим експлікантом змістовної структури тексту.

Відмінна риса розподілу тексту на мікроструктури є те, що такий поділ не обов'язково повинен збігатися з розподілом на абзаци. Присутня також й розбіжність при поділі на мікроструктури англійських і українських, ідентичних за змістом, текстів.

Наступним порушенням принципу регулярного коливання виступають речення, які починаються з визначених елементів (підвищення ентропії на початку речення). Такі речення в тексті можуть бути також представлені цілими групами. Вони виступають мікротемами у загальній структурі тексту, де всі наступні речення виступають їхнім розгорнутим описом. На основі таких речень можна скласти коротке уявлення суті оповідання.

Таким чином, на підставі дистрибутивного подання синтактико-семантичної структури тексту можна зробити наступні висновки:

- синтактико-семантичне подання тексту в термінах моделі ДССК здатне виявляти змістовну структуру тексту;
- структура тексту ділиться на тематичні мікроструктури;
- дроблення тексту на мікроструктури є регулярним;
- границі мікроструктур не завжди збігаються з розподілом тексту на абзаци;
- розподіл тексту на абзаци несе в собі структурно-логічний характер;
- усередині текстової макроструктури простежується чіткий тема-рематичний зв'язок на
- основі речень, що починаються з визначених елементів ДсмК;
- розподіл ідентичних за змістом, але різномовних тестів на мікроструктури не збігається.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ ТА ЛІТЕРАТУРИ

1. Тураева З.Я. Лингвистика текста. – М.; "Просвещение", 1986. – 126 с.
2. Новиков А.И. Семантика текста и её формализация. – М.; "Наука", 1983. – 215 с.
3. Шайкевич А.Я. Дистрибутивно-статистический анализ текстов: Дис. ... д-ра филол. наук: 10.02.21 – М.; 1982. ─ 32 с.
4. Мирам Г.Е. Дистрибутивная модель синтаксиса и семантики научно-технического текста для систем автоматической обработки информации: Дис. ... д-ра. филол. наук: 10.02.21 – М.; 1996. – 320 с.
5. Тер-Минасова С.Г. Словосочетание в научно-лингвистическом и дидактическом аспектах. – М.; 1981. – 175 с.

Матеріал надійшов до редакції 12.09.2006 г.

Морин С.В. Колебания речевой энтропии при анализе английских и украинских публицистических текстов в терминах дистрибуции.

В статье исследована структура текста путем анализа колебаний речевой энтропии на примере английских и украинских публицистических текстов с помощью дистрибутивной синтактико-семантической модели. Проведен сравнительный анализ структур указанных текстов. Методом дистрибуции доказано дробление текстовой структуры на микроблоки, а также наличие предложений, обладающих смысловой насыщенностью.

Morin S.V. Fluctuations of lexical entropy when analyzing English and Ukrainian publicistic texts in terms of co-occurrence.

In the article a textual structure has been investigated by means of analyzing lexical entropy fluctuations on the basis of English and Ukrainian publicistic texts. A comparative structural analysis of the texts under consideration has been

carried out. The division of a contextual structure into microblocks as well as the presence of contextually semantic sentences in it has been proved.