Орися Демсъка-Кульчицъка

БАЗОВІ ПОНЯТТЯ КОРПУСНОЇ ЛІНГВІСТИКИ

У структуру сучасних лінгвістичних дисциплін доволі активно вийшла відносно нова дисципліна - корпусна лінгвістика, до компетенції якої належить осмислення, вивчення й інтерпретація усіх процесів, пов'язаних з побудовою, обробленням, використанням і аналізом машиночитаного **корпусу** писемного і / або усного варіанта реалізації системи мови. Говоримо про певну відносність новизни цього напрямку, оскільки його початки сягають 60-х років минулого століття, точніше 1964 року, чи року створення першого корпусу природної мови - Brown Corpus або Браунівського корпусу [1], - повністю реалізованим в межах напрямку корпусної лінгвістики й одночасно електронним корпусом першого покоління.

Доречно зауважити, що корпусні методики дослідження мови далеко не продукт XX століття. Але лише після поєднання корпусних і комп'ютерних методик маємо підставу говорити про розвиток окремого напрямку, а не застосування методики, який особливо активно розвивається з початку 90-х років минулого століття в англістиці, а з середини 90-х - в усій романо-германській і слов'янській європеїстиці. Сьогодні ефективність корпусно-базованих досліджень тісно пов'язана з розвитком комп'ютерних наук.

В українській науці про мову корпусна лінгвістика як напрямок, особливо теоретичні аспекти цього напрямку та практичні застосування його положень, і, відповідно, термінологія мало відомі, тому в першій із статей, присвяченій цьому напрямку, передовсім розглянемо базові поняття та їх інтерпретацію: **корпус, природна мова, штучна мова, мова маркування, кодування, анотація, теґування, лематизація, слововживання, слововиділення, синтаксичний розбір, конкорданс.**

Корпус (corpus, мн. **corpora** або **corpuses),** як структуроване зібрання машиночитаних текстів найрепрезентативнішого мовного матеріалу тієї або іншої природної мови, є передовсім загальним методом фіксації та дослідження мови, узагальненою моделлю організації та подання фактичного матеріалу, базованим на

машинних технологіях. **Корпус** формується з реальних уривків писемного або усного мовлення, не передбачаючи модифікації мовленнєвої дійсності, що перетворює його на категорію емпіричну і дозволяє розглядати фактичний корпусний матеріал як емпіричну базу лінгвістичного дослідження. Під корпусом, крім машиночитаного структурованого зібрання текстів природної мови, інколи розуміють (1) будь-який текстовий матеріал і (2) довільний машиночитаний текст, хоча таке розуміння є дещо некоректним.

У корпусних дослідженнях чітко розмежовують поняття природної і штучної мови, оскільки обидва ці типи мов функціонують паралельно. Поняття **природної мови (natural language)** в корпусному мовознавстві розширює свою семантику щодо загальнолінгвістичної і також означає імітовану машиною форму людської мови. Натомість під поняттям **штучна мова** розуміють передовсім мови програмування і формальну логіку **(computer language, machine language, programming language, formal logic),** призначені для оброблення природномовної інформації програмними засобами. На сьогодні вихідною в межах корпусної лінгвістики штучною мовою є **Стандартна узагальнена мова маркування (Standard Generalized Mark-up Language = SGML),** чи певна стандартна система розмічування електронного тексту, на якій базуються комп'ютерні програми оброблення корпусів.

Програмне оброблення будь-якого тексту в корпусно-базованих роботах передбачає наявність **розмітки (mark-up),** тобто формалізовано поданої відповідної текстової та лінгвістичної інформації. Подання текстових і мовних даних у встановленому форматі називається **кодуванням (encoding),** однозначність застосувань якого до оброблення різних мов визначається міжнародним **Стандартом кодування корпусу (Corpus Encoding Standard = CES).** Кодування безпосередньо пов'язане з поняттям **анотації (annotation),** яке інтерпретується, по-перше, як практика додавання визначеної лінгвістичної інформації до машиночитаного тексту, по-друге, - як наявність цієї інформації у тексті, і, по-третє, - як сама така інформація. Залежно від типу анотованої інформації розрізняють:

— **анафоричну анотацію (anaphoric annotation) -** тип анотації, який передбачає наявність займенникових посилань у корпусі;

-дискурсну анотацію (discoursal annotation) - тип анотації, яка передовсім маркує елементи організації дискурсу.Через неоднозначність ідентифікації дискурсної інформації в тексті, цей тип анотації непоширений у корпусно-базованих дослідженнях;

-просодичну анотацію (prosodic annotation) - тип анотації, який охоплює суперсегментний рівень мовлення, передовсім наголос, інтонацію і паузи;

-семантичну анотацію (semantic annotation) - тип анотації, який передбачає маркування відношень між семантичними елементами тексту, наприклад розмітка компонентів сюжету або специфіки розгортання сюжету;

-фонетичну анотацію (phonetic transcription) - тип анотації, яка подає фонетичну специфіку усного мовлення природної мови і має формат фонетичної транскрипції.Повністю фонетично анотованих корпусів існує небагато, більшість фонетичних анотацій міститься в межах **просодичної анотації.**

(morphosyntactic annotation, part-of-speech annotation) - основний і найпоширеніший тип анотації у корпусних дослідженнях, який передбачає маркування частиномовної належності та частиномовних характеристик одиниць лексичного рівня.

Частиномовне анотування, чи **теґування (tagging) передбачає приписування теґів (tags,** одн. **tag)** - спеціально створених коґдів, за допомогою яких формалізується і задається відповідна морфологічна інформація про конкретне слово, до якого цей код приписано. Наприклад, *весна_Nfsn,* де *N* = іменник, *f* = жіночий рід, *s* = однина і *n* = називний відмінок, тобто код *Nfsn* несе інформацію про те, що лексема *весна* - це іменник жіночого роду в формі однини і називного відмінка; *говорити_Vin: V* = дієслово, *i* = форма інфінітива, *n* = недоконаний вид, отже, *говорити* - дієслово, інфінітив, недоконаного виду; *добрий_Amqsn: A* = прикметник, *m* = чоловічий рід, *q* = якісний, *s* = однина і *n* = називний відмінок, відповідно *добрий* - це прикметник чоловічого роду, якісний, у формі однини називного відмінка.

Залежно від функціональної специфіки, розрізняють звичайні або одиничні, дубльовані та гібридні теґи. **Дубльовані теґи (ditto tag) -** це, по-суті, той самий код, приписаний до кожного окремого елемента ідіоми. Через дубльований теґ зберігається ідіоматична єдність між елементами. **Гібридні теґи (portmanteau tag),** тобто паралельно приписані різні коди до одного мовного елемента, який може мати різні лінгвістичні характеристґики за умови збереження тих самих формальних ознак, наприклад прикметник і субстантивований іменник в українській мові. У корпусно-базованих дослідженнях теґ окремо, як правило, не функціонує, тут використовуються **базовий набір теґів (base tagset),** або просто **набір теґів (tagset),** де базовий набір - спеціальний набір теґів, які детермінують базову структуру елементів документа, в межах якого вони використовуються, а набір теґів - сукупність теґів, застосованих для анотування конкретного корпусу або тексту. Наприклад, в межах проекту **Ініціативи кодування тексту** (TEI) розроблено вісім базових наборів теґів для: (1) прозових текстів; (2) поезії; (3) драми; (4) транскрибованого усного мовлення; (5) листів і меморандумів; (6) словникових статей; (7) термінологічних статей; (8) корпусів і фондів.

Отже, коли йдеться про частиномовну / морфолого-синтаксичну анотацію, в лінгвістиці корпусу функціонує поняття **частиномовне теґування (part-of-speech tagging,** або **POS tagging).** Під частиномовним теґуванням також розуміють комп'ютерний інструментарій, призначення якого приписувати формалізовані частиномовні атрибути до слів у корпусі.

Стосовно слова, як елементарної одиниці корпусу, то, крім розмітки, над ним здійснюється так зване лематизаційне опрацювання, тобто на базі словоформи / словоформ виводиться лексикографічна або вихідна форма відповідного слова. Термін лематизація (lemmatisation) використовують на позначення процесу перетворення словоформ на відповідні лексикографічні форми лексем у корпусі. Відповідно лема (lemma) **- це вихідна форма слова, а лематизатор (lemmatizer) - програма, яка здійснює лематизацію в машинному середовищі. Крім понять лексикографічна форма слова / лема, словоформа / словоформи, у корпусному мовознавстві функціонує поняття токена (token), яке перекладають як лексема,** і йдеться про кожну конкретну словесну одиницю, з будь-якою квантитативною характеристикою, у тексті. Наприклад, у наведеному нижче текстовому уривку лемою є я, а слововживанням леми *я: я (2 рази), зі мною (2 рази).*

І тоді я оповім їй про те, що зі мною вчинив цей чаклунський вечір. Я став іншою людиною і не можу більше жити, як раніше. Хочу, щоб і ти збагнула, щоб пішла зі мною у небувале... (О. Бердник. Сузір'я зелених риб).

Слововживання об'єднуються в **тип (type) -** сукупності усіх форм і значень слова. Зауважимо, що інколи поняття лексема і слово чітко розмежовується сферою функціонування. Так, лексему відносять лише до сфери програмування, а слово - до лінгвістики, хоча такий розподіл є безпідставним. Корпусна лінгвістика - це напрямок, який вифор-мувався на стикові дисциплін, а отже, повинен оперувати як поняттями однієї, так і другої дисципліни, тобто лінгвістки і прикладної математики.

У корпусно-базованих дослідженнях, крім рівня слова, забезпечується рівень синтаксичної конструкції. Останній або об'єднується з рівнем слова і трактується як морфолого-синтаксичне теґування, або кожен з рівнів, не знищуючи логічних зв'язків між ними, реалізується індивідуально, і тоді йдеться про частиномовне теґування і синтаксичний розбір. **Синтаксичний розбір,** або **синтаксичний аналіз (parsing) -** це процес фіксації синтаксичної структури корпусного тексту, який здійснюється після ідентифікації базових морфолого-синтаксичних категорій відповідних елементів тексту. Є два типи синтаксичного розбору: **(1) повний синтаксичний розбір (full parsing)-** тип розбору, метою якого є забезпечення **якомога повнішого тадетальнішого аналізу речення;** і (2) **схематичний розбір (skeleton parsing), чи такий тип аналізу, який фактично стосується лише загальної структури синтаксичної одиниці. А синтаксично розмічений корпус (parsed corpus) - це корпус текстів природної мови, синтаксично проаналізований і відповідно розмічений.**

Метою маркаційного опрацювання тексту є наступний автоматичний пошук і отримання необхідної лінгвальної або лінгвістичної інформації для наукових досліджень, експериментів,

підтвердження теорії емпіричними даними. Одна із можливостей розміченого на рівні слова і синтаксичної одиниці корпусу є реалізація **конкордансу (concordance) -** словопокажчика, який зв'язує кожне слововживання у корпусі зі своїм контекстом. Конкорданс є домінуючою методикою корпусної лінгвістики, оскільки основне його призначенням - експлікація реальних текстових моделей. Конкордансна, чи більше поширений термін дистрибутивна методика, є відомим прийомом, який використовується у лінгвістиці уже доволі давно, але створювані машиною конкорданси гнучкіші та динамічніші, контекстне оточення в них може одночасно добиратися за багатьма критеріями, а візуалізація прикладів може відбуватися у будь-якому заданому порядку. В межах конкордансу виділяються поняття:

- **колокація (collocation) -** дефініюється як: (1) комбінація слів, які з високим ступенем ймовірності регулярно виступатимуть поряд у корпусі або тексті; (2) моделі появи слів у корпусі або тексті разом; і (3) моделі словосполучення.

Ідентифікація моделей співрозташування слів у текстових даних є важливим аспектом для укладання словників, вивчення мови й оброблення природної мови програмними засобами.

— **колокат (collocate) -** слово, яке в конкордансній моделі з'являється безпосередньо справа або зліва від центрального елемента, чи в термінології корпусної лінгвістики вузла (node), лінгвальна поведінка якого досліджується на предмет колокації;

—**контекст (context)** поняття, яке зберігає традиційне лінгвістичне значення, але у лінгвістиці корпусу йдеться про широкий контекст, натомість вузький контекст номінується як **спів-текст (co-text)** і йдеться про однослівну право- і лівобічну дистрибуцію визначеного слова, фраземи або елементарної синтаксичної конструкції.

Залежно від типу контексту - слово чи рядок - визначають різні типи конкордансу, які номінуються складними термінами **ключове слово** у **контексті (key word in context = KWIC),** де визначене слово візуалізується у спів-тексті, і **ключове слово і рядок (key word and line = KWAL) -** визначене слово візуалізується у кількарядковому контексті. Ключове слово і рядок є, по суті, лексикографічною карткою електронного типу.

— **рядок (string),** чи послідовна комбінація букв або набору символів;

— **інтервал (span) -** розмір спів-тексту або контексту, який визначають слова, тобто кількість двобічних елементів тесту до визначеного слова.

Прикладом українськомовного конкордансу, створеного в межах корпусної лінгвістики, є Конкорданції поетичних творів Тараса Шевченка [2].

Отже, у статті проаналізовано ряд базових понять нової лінгвістичної дисципліни - корпусної лінгвістики, які є або новими

для українського мовознавства, наприклад: **корпус, лематизація, мова маркування, анотація** тощо, або такими, що традиційно функціонують в мовознавчій науці, наприклад: **природна мова, штучна мова, слововживання, контекст, синтаксичний розбір** тощо. Подано лише базові поняття корпусного мовознавства, які, проте, репрезентують поняттєву картину цього сучасного лінгвістичного напрямку.

1. The Brown Standard Corpus of American English. – Brown University, 1964.
2. *Конкорданції поетичних творів Тараса Шевченка,* ред. О. Ільницький, Ю. Гавриш, Торонто: Наукове товариство Шевченка США, Канадський Інститут Українознавчих Студій, 2001.– т.1–4.

Мовна мозаїка

5(41). КОЛИ **ЗУСТРІЧАЮТЬСЯ** ДВОЄ Й КОЛИ **ТРАПЛЯЮТЬСЯ** ПОМИЛКИ

Надто часто мовці використовують дієслово *зустрічатися (зустрінутися, зустрітися)* там, де його має заступати дієслово *траплятися (трапитися).* Порушення літературних норм у цьому випадку зумовлене впливом російського слова *встречаться,* яке вживається порівняно з українським *зустрічатися* набагато ширше.

Українське *зустрічатися (зустрінутися, зустрітися)* стосується випадків, коли люди (рідше інші істоти) зіходяться, бачаться, бувають де-небудь разом або рухаються навпроти когось. Наприклад: ***Зустрілась*** *їй жінка, на плечах похилих несе щось, убога така* (Леся Українка); *Ми на вулиці* ***зустрілись***... (П. Тичина); *Сказать їй хочу так багато, а як* ***зустрінуся*** *– мовчу...* (В. Сосюра); ***Зустрілись*** *козаки й татари там, де могили над Дніпром...* (В. Сосюра); *До самого села не* ***зустрілося*** *жодної душі, тільки кам'яні хрести біліли над шляхом* (О. Гончар). Іноді замість особи виступають назви частин її тіла (руки, очі та ін.), наприклад: *У потиску* ***зустрілись*** *дружні руки* (М. Рильський); *Пархоменко звісив голову й* ***зустрівся*** *з колючими очима молодого чоловіка з тонкими губами* (П. Панч).

Дієслово *траплятися (трапитися)* поширюється на інші, ніж слово *зустрічатися*, ситуації. Здебільшого його вживають у значенні „відбуватися, діятися; бувати, бути, випадати", наприклад: *Так отак-то* ***Трапляється*** *в світі. Думав жити, поживати Та Бога хвалити, А довелось на чужині Тільки сльози лити!* (Т. Шевченко); *Вона була тільки тоді щаслива, як одпрошувалась в гості до батька, та й те* ***траплялось*** *дуже рідко* (І. Нечуй-Левицький); *Рідко* ***траплявся*** *такий щасливий рік, щоб не було пожежі в місті* (І. Нечуй-Левицький); *- В житті людини можуть* ***траплятися*** *такі випадки, що без болю їх згадувати не можна* (Г. Тютюнник). Дієслово *траплятися* передає також значення „несподівано з'явитися, прийти": *Тут самим нічого їсти, а, борони Боже,* ***трапиться*** *гість, тоді чим хоч приймай...* (М. Коцюбинський). І нарешті, наведемо ще значення „доводитися, мати нагоду". Тут форма третьої особи дієслова *траплятися* в ролі допоміжного засобу поєднується з неозначеною формою основного дієслова, наприклад: *Ніде мені не* ***траплялось бачить*** *такої ясної зеленої трави, такого густого зеленого листу на дереві, як у Карпатах* (І. Нечуй-Левицький).

Звідси випливає, що дієслово *зустрічатися* звичайно пов'язане з назвами людей, а *траплятися* – переважно з абстрактними іменниками.

Іван Вихованець