

АНАЛІЗ СИНТАКСИЧНОЇ СТРУКТУРИ УКРАЇНСЬКОГО РЕЧЕННЯ: КОРПУСНИЙ ПІДХІД

Чейлитко Наталія Геннадіївна,

канд. філол. наук

Київський національний університет імені Тараса Шевченка

Стаття присвячена проблемі вибору теоретичної моделі, яка стане основою синтаксично марковано-го корпусу української мови. Розглянуто принципи моделювання синтаксичної структури речення в термінах дерева залежностей; визначено недоліки вказаної моделі та запропоновано шляхи їх подолання.

Ключові слова: корпус, синтаксична структура речення, дерево залежностей, граф.

Формування корпусу текстів будь-якої мови має починатися зі з'ясування тих дослідницьких та / або навчальних завдань, які передбачається розв'язати на матеріалі створеного корпусу. Іншими словами, призначення корпусу визначає особливості його будови. Із наукового погляду корпус будь-якої мови необхідний насамперед для виявлення закономірностей функціонування мовних одиниць, що стає можливим лише у результаті послідовного аналізу великої кількості мовленнєвого матеріалу. Для української мови такого корпусу текстів досі не створено. Тому **актуальними** є ті наукові розвідки, які присвячені виробленню засад укладання корпусу української мови і – головне – їх апробації на українському мовленнєвому матеріалі.

Відповідно, дослідникові потрібно визначити репрезентативну вибірку текстів, які стануть основою майбутнього корпусу, оскільки цей фактор, а саме достатній обсяг вибірки, пропорційне представлення та однорідність підвибірок, належна па-спортизація текстів тощо, є обов'язковою умовою достовірності наукових висновків, які будуть отримані в результаті дослідження певного мовного явища на матеріалі сформованого корпусу.

Поставлені пошукові завдання зумовлюють вибір лінгвістичних моделей, за допомогою яких у корпусі описуються мовні явища: фонетичні, морфологічні, синтаксичні, семантичні. Очевидно, модель, обрана науковцем для характеристики елементів тексту, має бути максимально формалізованою, спроможною пояснити всі можливі, навіть нетипові, мовленнєві прояви мовних одиниць. Тому корпус можна вважати важливим засобом верифікації різноманітних теорій.

Мета цієї статті полягає у представленні моделі, вибраної для формального опису синтаксичної будови мовленнєвих одиниць у складі корпусу української мови. **Об'єкт** дослідження становить синтаксична структура українського речення, а **предмет** – її формальний опис. **Новою** є запропонована методика відображення у корпусі української мови синтаксичної структури речення як графа.

У вітчизняній лінгвістиці склалася традиція представляти синтаксичну організацію речення за допомогою орієнтованого дерева залежностей (ДЗ), вузли якого відповідають словоформам речення, а гілки (ребра) марковані відповідно до типу синтаксичного зв'язку між словоформами. Таким чином, ДЗ відображає структуру речення як набір бінарних зв'язків підпорядкування між словоформами. Причому

словоформою вважають послідовність буквених символів між двома пробілами (за іншою термінологією – це машинне слово або текстоформа). Побудова ДЗ відповідає таким умовам:

1. Дерево повинно мати тільки одну вершину (рис. 1).
2. Кожній словоформі у реченні відповідає один вузол у ДЗ.
3. Ребро відповідає зв'язкові між двома словоформами. Напрямок ребра – від словоформи, яка підпорядковує ("хазяїн"), до словоформи, яка підпорядковується ("слуга").
4. Усі ребра є орієнтованими, тобто напрям ребер дерева наперед визначений: від вершини до вузлів, які розташовані в дереві рівнем нижче.
5. В один вузол може входити не більше одного ребра.
6. Дерево має бути зв'язним, тобто у дереві не може бути вузлів, не пов'язаних з іншими вузлами дерева.
7. У дереві не може бути циклів, тобто ребро не може входити у вузол, із якого воно виходить.
8. Вершиною дерева можуть виступати присудки простого двоскладного речення, головний член односкладного речення, присудок головної частини складнопідрядного речення.

Така інтерпретація синтаксичної структури речення ґрунтується на лінгвістичній моделі "Смисл $\Leftrightarrow$ Текст" І.О. Мельчука [Мельчук 1974].

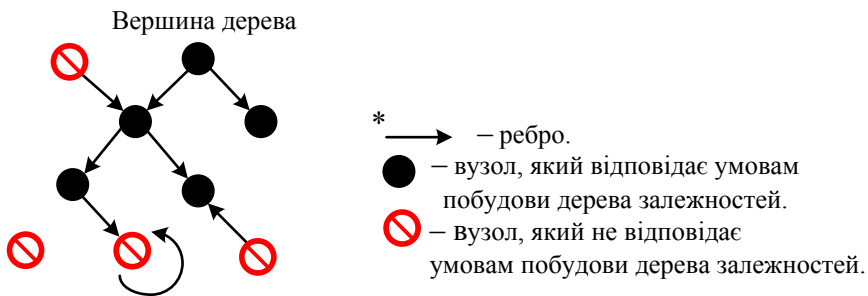


Рис. 1. Узагальнений вигляд дерева залежностей*

Досвід застосування дерева залежностей для корпусного представлення синтаксичної організації речення демонструє глибоко анотований корпус російської мови [Апресян 2005], створений на матеріалі сучасної російської прози, науково-популярних і суспільно-політичних статей та Інтернет-новин інформаційних агенцій (близько 20 тис. речень).

Нами було організовано експериментальний корпус текстів українських Інтернет-новин (близько 3 тис. речень), синтаксично розмічений у термінах дерева залежностей. На основі цього корпусу вдалося встановити закономірності реалізації синтаксичних зв'язків словоформ в українському реченні [Чейлитко 2009]. Набутий досвід

засвідчив доцільність репрезентації синтаксичної будови українського речення в термінах дерев залежностей і водночас указав на певні недоліки такого представлення.

1. Вузлом у дереві позначається словоформа як машинне слово (послідовність буквених символів від пробілу до пробілу), через що складові частини аналітичної лексики розглядаються окремо (*Ірина*→*Марківна*→*Попова*). Щоб показати належність усіх складників до однієї лексики, необхідно спеціально маркувати ребра між вузлами, які відповідають цим частинам.

2. У дереві залежностей неможливо відобразити зв'язок між словоформою та словосполученням (*єдиний* ← (*зважений коментар*)).

3. Вимога зв'язності дерева, тобто відсутності вузлів, в які не входило б жодного ребра, змушує дослідника ігнорувати випадок, коли у реченні наявний еліпсис і, отже, має місце розрив синтаксичних зв'язків (*Єдиний зважений та об'єктивний коментар* – [еліпсис] *Ірини Марківни Попової*). Для того, щоб уникнути такого розриву, доводиться штучно запроваджувати нульовий вузол, який позначає невиражену словоформу, і тим забезпечити зв'язність усіх вузлів дерева (рис. 2 А).

4. Позначення у дереві лише підрядного зв'язку між словоформою-хазяїном і словоформою-слугою за допомогою орієнтованих вертикальних ребер не дає можливості адекватно показати сурядний – горизонтальний – зв'язок між рівноправними, незалежними одна від одної словоформами (*зважений–та–об'єктивний коментар*). Для відображення сурядного зв'язку між словоформами використовують спеціально запроваджений уявний вузол: сурядні словоформи підпорядковуються у дереві уявному вузлові (на рис. 2 А від уявного вузла залежать сурядні словоформи *зважений*, *об'єктивний* і сполучник сурядності *та*).

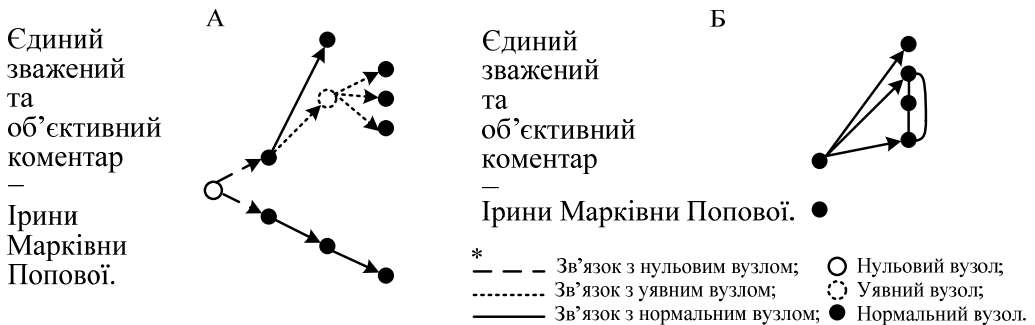
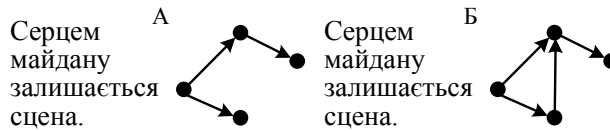


Рис. 2. Дерево залежностей для речення з аналітичною лексемою, сурядністю та еліпсисом*.

Структуру речення представлено:

А – із залученням додаткових вузлів, Б – без залучення додаткових вузлів

5. Правило, за яким в один вузол входить два і більше ребер, унеможливилює відображення зв'язків із дуплексивом – елементом речення, який підпорядковується двом членам речення (рис. 3: *Серцем майдану залишається сцена*).



**Рис. 3. Дерево залежностей для речення з дуплексивом.
Подвійний зв'язок: А – не відображено, Б – відображено**

Окрім цього, така ситуація позбавляє можливості показати зв'язки двох і більше сурядних словоформ з їхньою спільною залежною словоформою: *виваженість і толерантність людей* – словоформа *людей* одночасно підпорядковується двом сурядним словоформам *виваженість*, *толерантність*; відповідно, у вузол, який відповідає цій словоформі, мало б входити одразу два ребра (рис. 4). За спеціальною домовленістю спільний залежний компонент може підпорядковуватися у дереві уявному вузлові, який, окрім того, підпорядковує собі сурядні вузли (на рис. 4 А від уявного вузла залежить словоформа *людей*, а на рис. 5 А – словоформа *учасники*).

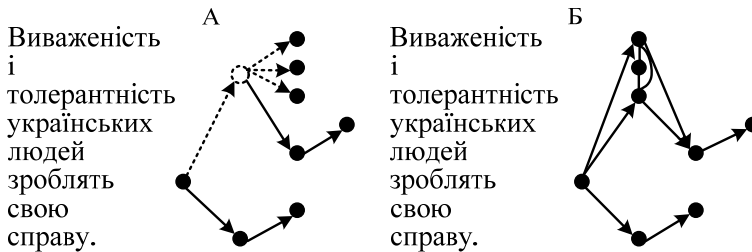
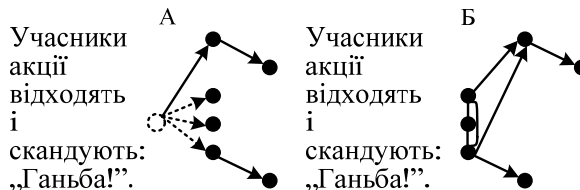


Рис. 4. Представлення зв'язку між сурядними словоформами та їхньою спільною залежною словоформою за допомогою уявного вузла (А) та без нього (Б)

6. Правило, за яким у дереві не може бути більше однієї вершини, спричиняє ситуацію, коли стає неможливим, без запровадження штучного уявного вузла, показати сурядність двох присудків: *учасники відходять і скандують: „Ганьба!”* (рис. 5).



**Рис. 5. Відображення сурядності присудків
за допомогою уявного вузла (А) та без нього (Б)**

Таким чином, моделювання синтаксичної організації речення із застосуванням дерева залежностей наражається на традиційні проблеми синтаксису, що ускладнює адекватне представлення цих синтаксичних ситуацій.

Уведення до дерева службових вузлів (нульового та уявного), і, відповідно, вироблення додаткових правил (домовленостей) щодо побудови дерева, ускладнює як саму методику, так і структуру дерев, створених на основі цієї методики. Так, зокрема, запровадження уявного та нульового вузлів може спричинити збільшення рівнів у дереві: на рис. 2 А – висота дерева становить 4 рівні, а на рис. 2 Б – всього 2 рівні. Зважаючи на те, що висота є однією із найважливіших характеристик дерева, на яку спираються у процесі стилістичної діагностики текстів [Севбо 1981], штучне зростання цього параметра є укріпленням небажаним.

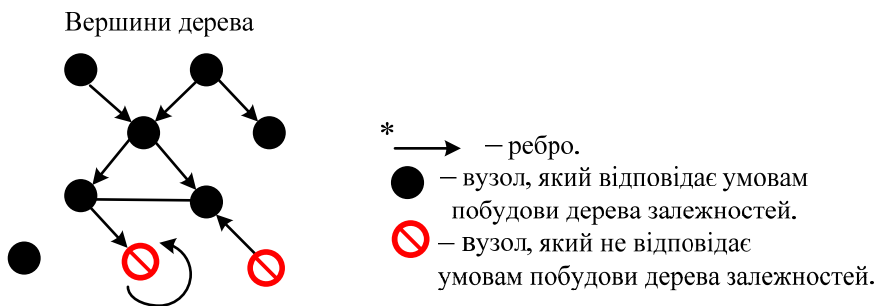


Рис. 6. Узагальнений вигляд графа для відображення синтаксичної структури речення*

На нашу думку, моделювання синтаксичної структури речення необхідно здійснювати за допомогою більш загальної моделі – графа, по відношенню до якого орієнтоване дерево залежностей є його підвидом – рис. 6). Це дасть можливість:

- ⇒ відображати розрив синтаксичних зв'язків у реченні з еліпсисом через неприємні вузли графа;
- ⇒ позначати горизонтальними ребрами сурядний зв'язок між словоформами;
- ⇒ показувати подвійний синтаксичний зв'язок (дулексив), запровадивши дозвіл на входження двох ребер в один вузол;
- ⇒ визначати дві або більше вершини графа, які вказуватимуть на наявність сурядних присудків у реченні;

Репрезентація структури речення у вигляді графа більшою мірою відповідає природі реального речення. Окрім того, таке представлення значно спростить процедуру побудови графічної моделі речення, зробить її більш відкритою до автоматичного обчислення її якісних та кількісних параметрів.

Статья посвящена проблеме выбора теоретической модели, которая станет основой синтаксически маркированного корпуса украинского языка. Рассмотрены принципы моделирования синтаксической структуры предложения в терминах дерева зависимостей; определены недостатки указанной модели и предложены пути их преодоления.

Ключевые слова: корпус, синтаксическая структура предложения, дерево зависимостей, граф.

The article is devoted to the problem of selecting a theoretical model for syntactically annotated corpora of the Ukrainian language. Principles of modeling syntactical sentence structure in terms of dependency tree are showed. The imperfections of the model are identified and ways to overcome them are proposed.

Keywords: *corpora, syntactic structure of sentences, dependency tree, graph.*

Література:

1. Апресян Ю.Д. Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы / Ю.Д. Апресян, И.М. Богуславский, Б.Л. Иомдин и др. // Национальный корпус русского языка: 2003–2005. – М.: Индрик, 2005. – С. 193–214.
2. Мельчук И.А. Опыт теории лингвистических моделей "Смысл-Текст" / И.А. Мельчук. – М.: Наука, 1974. – 314 с.
3. Севбо И.П. Графическое представление синтаксических структур и стилистическая диагностика / И.П. Севбо. – К.: Наукова думка, 1981. – 192 с.
4. Чейлитко Н. Корпусне дослідження зон зв'язків словоформ в українському реченні / Н. Чейлитко // Лінгвістичні студії : [зб. наук. праць]. – Донецьк, Вид-во ДонНУ, 2009. – Вип. 18. – С. 268–275.