Ганна Ситар
(м. Вінниця)

СИНТАКСИЧНА ІНДЕКСАЦІЯ КОРПУСУ УКРАЇНСЬКОЇ МОВИ: ПРОБЛЕМИ І ПЕРСПЕКТИВИ

Важливою рисою лінгвістичних досліджень у XXI столітті є корпуснозорієнтованість, що висуває на перший план потребу створення великих корпусних проектів, виконаних на матеріалі різних мов. В Україні одним із великих і доступних для широкого загалу корпусних проектів є Корпус української мови, створений під керівництвом Наталії Дарчук колективом лабораторії комп'ютерної лінгвістики Інституту філології Київського національного університету імені Тараса Шевченка і розміщений за адресою http://www.mova.info/corpus.aspx.

Теоретичним підґрунтям індексування текстів у цьому проекті є «комп'ютерна граматика АГАТ із вбудованими процесорами опрацювання українськомовного тексту» [2, с. 148] (докладно про граматику АГАТ див. у монографії Наталії Дарчук [3]).

Здійснюваний у Корпусі української мови автоматичний синтаксичний аналіз грунтується на двох принципах:

1) аналіз речення за **безпосередніми складниками**, ідея якого належить Леонарду Блумфілду;

2) побудова для речення **дерева залежностей**, правила якої визначив Люсьєн Теньєр.

Вибір цих принципів є невипадковим: обидва підходи запропоновано в межах структуралізму, відповідно вони втілюють суто формальний підхід до речення і виявляються найпридатнішими для виконання автоматичного аналізу. Водночас накопичений на сьогодні досвід синтаксичного розмічення корпусів англійської, шведської, російської та інших мов базується саме на принципах граматики залежностей і передбачає побудову дерев залежностей з визначенням головного й залежного компонентів і встановленням типу синтаксичного зв'язку між ними і/або синтаксичного відношення [6; 7; 8; 9; 5 та ін.].

На етапі експертної перевірки результатів автоматичного синтаксичного аналізу до проекту долучились викладачі кафедри загального та прикладного мовознавства і слов'янської філології та студенти спеціальності 035 «Філологія» освітньої програми «Прикладна лінгвістика» Донецького національного університету імені Василя Стуса.

У процесі корегування результатів автоматичного поділу речень на двокомпонентні комплекси й визначення залежностей між парами словоформ було виявлено окремі моменти, які, очевидно, потребують подальшого обговорення.

Зокрема, у межах кваліфікації безпосередніх складників велику увагу приділено визначенню морфологічного статусу головного і залежного слів, що, безумовно, є цілком виваженим, оскільки:

а) автоматичний морфологічний аналіз є необхідною передумовою для виконання автоматичного синтаксичного аналізу;

б) одним із критеріїв (але не єдиним) класифікації словосполучень є саме частиномовна належність їх компонентів.

Водночас основними категоріями формально-синтаксичного аналізу є **синтаксичний зв'язок** і **член речення** [1; 4 та ін.], кваліфікація яких і має бути завданнями здійснюваного синтаксичного аналізу. Так, для речення *Сьогодні учні виконали складне завдання* комп'ютерна програма приписує однакові параметри парам слів *виконали сьогодні* і *виконали завдання*, оскільки за прийнятою у проекті класифікацією з погляду частиномовного статусу обидва словосполучення є дієслівними безприйменниковими, хоча з погляду формально-синтаксичного вони поєднані різними типами синтаксичного зв'язку – відповідно приляганням і керуванням. Тому вважаємо за доцільне розмежувати кілька параметрів аналізу для виділених пар слів:

1) частиномовна належність головного і залежного компонентів;

2) тип синтаксичного зв'язку, що їх поєднує (розмежування словосполученнєвотвірних (керування, прилягання (власне-прилягання і відмінкове прилягання), узгодження), реченнєвотвірних (предикативний на рівні простого, сурядний і підрядний зв'язки на рівні складного речення (із подальшою диференціацією підтипів)) і реченнєвомодифікувальних (сурядний, опосередкований, напівпредикативний та ін.) зв'язків обґрунтовано у працях Анатолія Загнітка [4 та ін.]);

3) член речення (підмет, присудок, додаток, обставина, означення, прикладка) / компонент у складі відповідного члена речення (за умови семантичної цілісності типу *повинен був працювати, батько з матір'ю, двадцять дві книжки*) / не є членом речення (наприклад, підрядний сполучник у складнопідрядному реченні).

У перспективі таке розмежування дасть змогу встановити закономірності співвідношення / неспіввідношення типу синтаксичного зв'язку, морфологічного наповнення компонентів та членів речення, якими вони виступають. Важливим також вважаємо передбачити варіант «інше» в межах усіх параметрів, оскільки, як свідчить практика, спрогнозувати всі можливі вияви синтаксичної реалізації є проблематичним. Наприклад, у реченні *У Державній податковій адміністрації України (ДПАУ) сьогодні і так удосталь повноважень для нормального збору податків* абревіатура ДПАУ залишилась поза межами комп'ютерного аналізу.

Нечастотними, проте такими, які потребують уваги програміста, виявились помилки у встановленні меж речення. У таких випадках йшлося

про складні багатокомпонентні конструкції з різними видами зв'язку та прямою мовою. Зрозуміло, що в класичних працях, присвячених граматиці залежностей, подібні речення не розглядалися. Проте аналізовані в межах публіцистичного підкорпусу тексти газети «День» продемонстрували випадки штучного розривання частин речення, очевидно, через складність пунктуаційного оформлення. Так, роз'єднання підрядної і головної частин унеможливлює визначення зв'язків між ними та правильну побудову дерева залежностей, наприклад:

Коли спитати громадянина України: «Хто Ви?

і

», він відповідає: «Я українець російського походження», або «Я українець єврейського походження», або «Я українець гагаузького походження», або «Я українець із діда-прадіда»[1].

Тому пропонуємо передбачити можливість для лінгвіста-експерта об'єднати два чи більше речення в одне за допомогою введення спеціальної опції.

Наступним етапом досліджень є автоматичне визначення синтаксичних (семантико-синтаксичних) відношень у межах виділених складників, що можуть бути відображені в межах побудованих дерев залежності.

Література

1. Вихованець І. Р. Граматика української мови. Синтаксис. Київ: Либідь, 1993. 368 с.

2. Дарчук Наталія. До питання про створення автоматичного словника словосполучень української мови. *Лінгвістичні студії / Linguistic Studies:* зб. наук. праць. Вінниця, 2018. Вип. 36. С. 148−158.

3. Дарчук Н. П. Комп'ютерне анотування українського тексту: результати і перспективи / монографія. Київ: Освіта України, 2013. 544 с.

4. Загнітко А. П. Теоретична граматика української мови: Синтаксис: Монографія. Донецьк: ДонНУ, 2001. 662 с.

5. Национальный корпус русского языка. URL: http://www.ruscorpora.ru (дата звернення 10.03.2019).

6. Nivre Joakim. Introduction to Dependency Grammar and Dependency Parsing (Uppsala University). URL: ufal.mff.cuni.cz/~bejcek/parseme/prague/Nivre1.pdf (дата звернення 10.03.2019).

7. Nivre Joakim. Inductive Dependency Parsing. Springer, 2006. 212 p.

8. Silveira Natalia, Dozat Timothy, de Marneffe Marie-Catherine, Bowman Samuel R., Connor Miriam, Bauer John and Manning Christopher D. Gold

[1] Зверніть увагу на пунктуаційне оформлення частин речення, виділених програмою як окремі речення.

Standard Dependency Corpus for English. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014).* 2014. P. 2897–2904. URL: https://aclweb.org/anthology/papers/L/L14/L14-1067/ (дата звернення 10.03.2019).

9. Universal Dependencies. URL: https://universaldependencies.org/ (дата звернення 10.03.2019).

Микита Яблочков
(м. Київ)

СТВОРЕННЯ КОМП'ЮТЕРНОГО ІНСТРУМЕНТАРІЮ ДЛЯ ДОСЛІДЖЕННЯ СТРУКТУРИ СТАТТІ ТЛУМАЧНОГО СЛОВНИКА: НА ПРИКЛАДІ СЛОВНИКА ІСПАНСЬКОЇ МОВИ

На основі аналізу інтерфейсів цифрових тлумачних словників (СУМ-20 – http:// http://services.ulif.org.ua/expl, Оксфордський словник англійської мови – http://en.oxforddictionaries.com/, Словник іспанської мови Королівської академії іспанської мови – http://dle.rae.es), що відрізняються структурами і методами укладання, виникла ідея створити інструмент для більш глибокого аналізу їх структур в одній концептуальній схемі. В цій статті описані перші етапи цієї роботи.

Цей інструментарій створюється як для аналізу структури статей тлумачного словника, так і усього словника в цілому, перевірки прозорості лексикографічних даних словника (виявлення імпліцитної інформації, структур та закономірностей), а також для можливості швидкої зміни зовнішніх інтерфейсів.

Як матеріал дослідження було вирішено використати текст Словника іспанської мови Королівської академії іспанської мови (http://dle.rae.es). Цей вибір був зумовлений наступним: принциповою відмінністю підходу до формування структури словника від прийнятої в УМІФ та тої, що використовується в Оксфордському словнику англійської мови (Oxford Dictionary), наявністю проведених досліджень лексикографічної системи іспанської мови [2], а також наявність цифрової версії словника у форматі HTML5, що гарантує автентичність тексту (його лінгвістичну достовірність) і дозволяє повністю зосередитися на його структурі.

Крім цього важливою особливістю цього словника є високий рівень складності структури словникових статей та велика кількість лінгвістичної інформації (включаючи граматичну, етимологічну та іншу), що включена в усі статті словника. Ця лексикографічна праця містить основну частину