УДК 81’322

Наталія Дарчук
Київський національний університет імені Тараса Шевченка

АВТОМАТИЧНИЙ СИНТАКСИЧНИЙ АНАЛІЗ ТЕКСТІВ КОРПУСУ УКРАЇНСЬКОЇ МОВИ

Розглядається автоматичне представлення синтаксичної структури речення на рівні словосполучення: автоматичне виокремлення словосполучення, приписування йому типу синтаксичного зв’язку (підрядного, сурядного, предикативного).

Ключові слова: автоматичний синтаксичний аналіз, словосполучення, синтаксичний зв’язок, підрядний зв’язок, сурядний зв’язок, предикативний зв’язок, ядровий підрядний зв’язок, ад’юнктний підрядний зв’язок.

Автоматичний синтаксичний аналіз (АСА) – проект, над вирішенням якого працюють розробники Корпусу української мови, співробітники лабораторії комп’ютерної лінгвістики Інституту філології Київського національного університету імені Тараса Шевченка.

На рівні словосполучення АСА передбачає автоматичне виокремлення словосполучень, приписування їм типу синтаксичного зв’язку та автоматичне укладання словників словосполучень (дієслівних, іменних, ад’єктивних). На рівні речення має здійснюватися повний синтаксичний аналіз у вигляді дерев залежностей.

За результатами роботи над проектом передбачається створити електронний алфавітно-частотний словник сполучуваностей української мови, який по завершенні дослідницької роботи буде викладений в Інтернеті на мовному порталі www.mova.info для загального користування.

Для української мови подібне лінгвістичне та програмне забезпечення розробляється вперше, тобто цей проект має унікальний характер. Це єдиний лінгвістичний ресурс, що містить синтаксичне розмічування текстів української мови, яке здійснюється тільки автоматично на базі повного автоматичного морфологічного аналізу зі знятою омонімією (про перші спроби АСА див. [Дарчук 1990; Дарчук 1999]).

У теоретичному плані вирізнення словосполучення з реченнєвої структури на великих різностильових масивах текстів, які входять до Корпусу української мови, дає можливість дослідникам української мови визначити синтаксичну і семантичну ємність цієї синтаксичної одиниці.

Необхідність вивчення сполучуваності лексичних одиниць зумовлена нерозробленістю широкого кола як теоретичних, так і прикладних проблем. Теоретичні аспекти, які потребують вивчення, – це, зокрема, граматична і лексична валентність слів, типова сполучуваність, синонімія словосполучень різних структурних типів, лексична і граматична валентність як критерій синонімічності, закони комбінаторики словосполучень різних типів і розрядів, лексична валентність як критерій розмежування вільних і фразеологічних словосполучень, взаємодія стійкості й ідіоматичності тощо. До прикладних проблем можна віднести автоматизацію лінгвістичних досліджень, автоматичне визначення меж словосполучень, установлення критеріїв членування фрази на синтагми, автоматичний синтаксичний аналіз речення, автоматичне реферування й анотування тексту на основі сполучувальнісних критеріїв тощо.

Базою для розроблення, впровадження і застосування АСА є Корпус української мови, який постійно зростає. На початок листопада 2012 р. його обсяг сягав близько 13 млн. слововживань, або близько 650 тис. речень. Отже, стала потреба у створенні потужного механізму автоматичного опрацювання українського тексту на рівні синтаксису і, відповідно, у розробленні лінгвістичного та програмного забезпечення цього ресурсу. Постали завдання створення такого типу АСА, за допомогою якого можна одержати різноманітну інформацію про функціонування граматичних синтаксичних одиниць та їх категорій. При цьому виникає дилема обсягу матеріалу і точності його опрацювання. Створення аналізатора, який би абсолютно безпомилково здійснював аналіз українського тексту, неможливе, тому якісне анотування тексту завжди пов’язане з ручним доопрацюванням.

У цьому сенсі за умови відносної обмеженості організаційних можливостей розробники Корпусу опинилися перед вибором: створення порівняно невеликого, але вивіреного корпусу чи значного за обсягом, але анотованого автоматично. Обидва підходи мають право на існування. Ми обрали другий з них: розроблення лінгвістичного і програмного забезпечення, за допомогою якого з будь-якого тексту Корпусу автоматично виділяються словосполучення з подальшою можливістю редагування одержаних даних. На цьому матеріалі, так само в автоматичному режимі, будуються словники сполучуваності для різних частин мови: окремо для слів, що виступають у ролі ядрових (“хазяїн”) і в ролі ад’юнктних (“слуга”).

Не вдаючись до теоретичних дискусій щодо деяких питань синтаксису, зазначимо, що в основі АСА лежить формально-синтаксичний аспект вивчення речення. Ані семантико-синтаксичний і функціональний, ані комунікативний підхід до розгляду речення не можуть стати основою автоматизації. Тоді як дослідження формально-синтаксичної будови речення дає можливість створити словник синтаксем, для якого попередньо слід укласти таксономічну класифікацію лексики, що у майбутньому уможливить автоматичне визначення синтаксичних відношень між членами словосполучення. Формальна граматика, адаптована для потреб автоматизації, базуватиметься на гіпотаксисі як провідному аспекті синтаксичного ладу мови; а паратаксис буде додатковим аспектом, оскільки виокремлення сурядних словосполучень з погляду автоматизації не становить суттєвих труднощів.

У ході автоматичного синтаксичного аналізу речення насамперед має здійснюватися автоматичний пошук зв’язків слів у реченні. Ознаки таких зв’язків наявні, зокрема, у словозмінних характеристиках слів. У реченні послідовно розгортається підпорядкування слів одне одному: одне слово (залежне) змінює форму, щоб адаптуватися до вимог іншого слова (головного).

Таким чином, машина має виокремлювати пари слів, пов’язані граматичним зв’язком, позначаючи напрямок залежності.

Наприклад, для речення: *Широко* (1) *обговорюються* (2) *проблеми* (3) *життя* (4) *українського* (5) *суспільства* (6). – виокремлюються такі пари слів:

(5) ← (6) українського суспільства
(1) ← (2) широко обговорюються
(3) → (4) проблеми життя
(4) → (6) життя суспільства
(2) ↔ (3) обговорюються проблеми

Цим діям можна надати алгоритмічного вигляду. Врешті отримуємо список пар залежностей. Жодних даних семантичного характеру у цьому аналізі не використовується. Єдине, що можна визначити при такому підході, – це залежність слів одне від одного і порядок їх розташування. Це і є прикладом формально-синтаксичного підходу до аналізу речення.

Словосполучення – це смислове та граматичне поєднання двох або більшої кількості слів на основі підрядного, сурядного або предикативного зв'язку [Загнітко]. Ці типи зв'язків відповідають відтворенню загальної системи відношень між компонентами описуваної ситуації у реченні. Віднесення до словосполучень тільки тих, які сполучаються підрядним прислівним зв'язком, не є вичерпним з точки зору складників речення [Вихованець].

Ми вважаємо, що словосполучення – відносно самостійна одиниця мови, що виділяється у межах речення, будується за законами поєднання слів, виявляє у мовленні валентні властивості головного слова, має мовні моделі, відтворювані у мовленні. Не є словосполученнями: складені аналітичні поєднання слів, зокрема сполуки іменника з прийменником (*через міст, в інституті*); складені аналітичні форми слів (*буду читати, більш досвідчений*); фразеологізми (*ні пари з вуст, бити байдики*).

Завданням АСА є виявлення всіх різновидів сполучуваності – предикативної, підрядної і сурядної – кожного слова з текстів. Граматичні характеристики словосполучення безпосередньо залежать від того, до якої частини мови належить слово-"хазяїн", тому що лексико-граматична природа слова визначає його здат-

ність сполучатися з іншими словами. Відповідно до цього словосполучення поділяють на іменникові, прикметникові, займенникові, числівникові, дієслівні та прислівникові. За виробленою концепцією АСА при виокремленні словосполучень було передбачено попередній етап створення **словника валентностей** для дієслова (31 206 правил), іменника (40 023), ад'єктива (6205), а також словника фразеологізмів (близько 3000 одиниць).

За складом словосполучення поділяють на прості, складні та комбіновані. Ми виділяємо тільки прості бінарні словосполучення, які можуть бути поширені у складні або комбіновані автоматизовано, оскільки при визначенні їх складу потрібен аналіз смислової структури.

Якщо у сполуках у головній позиції слово інформативно недостатнє, а залежне слово цю недостатність заповнює (так звані доповнювальні, або комплетивні відношення), то вони розглядаються як словосполучення, що виконують функцію одного члена речення, наприклад: *дехто з присутніх, четверо з них, почав працювати* і под.

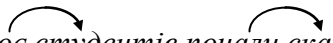
Ми відмовилися від традиційного поділу підрядного зв'язку на підвиди – узгодження, керування та прилягання. Ґрунтуючись на широкому розумінні поняття синтаксичного зв'язку, всі вони є випадками приєднання до головного слова відмінкової форми іменника або субстантива. При узгодженні залежне слово уподібнюється до головного в усіх його граматичних формах, а при приляганні воно, не маючи форм словозміни, приєднується до головного за змістом. Крім того, останнім часом у деяких роботах випадки поєднання з головним словом залежної форми іменника з атрибутивним чи обставинним значенням трактуються як прилягання [Русская граматика : 21]: *працювати лікарем* – зв'язок керування, а *прогулюватися парком* – прилягання; *допомога матері* – керування; *пам'ятник поетові* – прилягання.

Причиною такої відмови є ще й неможливість у деяких випадках автоматично, не вдаючись до аналізу значення кожного з членів словосполучення, визначити його тип. Алгоритм АСА має спиратися виключно на морфологічні форми слів (орудний відмінок залежного слова у першій парі; давальний у другій).

Підрядні зв'язки поділяються нами на ядрові і неядрові. Ядровим називаємо такий зв'язок, при якому аналізоване слово є керувальним, головним. Наприклад, у реченні – *Від економічної кризи сильно постраждали майже всі європейські держави*. – спостерігаємо такі ядрові зв'язки: ***кризи*** домінує над *економічної*; ***постраждали*** домінує над *від*; ***від*** домінує над *кризи*; ***постраждали*** домінує над *сильно;* ***всі*** домінує над *майже*; ***держави*** домінує над *всі*; ***держави*** домінує над *європейські*. Неядровий зв'язок – це такий зв'язок, при якому аналізоване слово є залежним, керованим. У попередньому прикладі неядрові зв'язки є у словах *економічної* (залежить від *кризи*), *сильно* (залежить від *постраждали*), *європейські* (залежить від *держави*) тощо.

Предикативний зв'язок – це зв'язок між основними компонентами речення "підмет – присудок", який ґрунтується на їхній двобічній залежності. У предикативній парі жодне зі слів не можна вважати домінованим, вони обидва є однаково домінувальними.

Якщо підмет або присудок виражений складеним словосполученням, то визначається підрядний зв'язок для аналізованого слова, наприклад:



Предикативний зв'язок установлюємо між *двоє студентів* і *почали скандувати*. В межах словосполучення *двоє студентів* ядровим буде ***двоє***, яке домінує над *студентів* (*студентів*, відповідно, має неядровий зв'язок); у словосполученні ***почали*** *скандувати* ядровим буде ***почали***, яке домінує над *скандувати*. Те саме стосується іменного складеного присудка:

Ядровий зв'язок встановлюємо між підметом *він* і присудком *став студентом*. У межах іменного складеного присудка ***став*** *студентом* ядровим буде допоміжне дієслово ***став***, тому що

воно домінує над іменною частиною, вираженою іменником в орудному відмінку *студентом*.

Сурядний зв'язок – це зв'язок, при якому жодне із взаємопов'язаних слів не є ані домінувальним, ані домінованим. Вважається, що два слова знаходяться в сурядному зв'язку, якщо кожне з них підпорядковане одному й тому ж третьому слову, якщо вони пов'язані через сполучник між собою або відокремлені одне від одного комою. При цьому ми дотримуємося таких умов, що сурядний зв'язок установлюється між словами, а не між словом і зворотом або синтаксичною конструкцією. Наприклад, сурядний зв'язок є між словами *чесними і прозорими* (*Вибори були чесними і прозорими*) і його немає у такому прикладі: *Президент був задоволений і в гуморі*.

Щодо ад'єктивів (прикметників, дієприкметників, займенників), які виконують одну й ту саму функцію, прийняті такі домовленості:

- якщо між ними є сполучник, то напрямок зв'язку такий: від головного слова до кожного з прикметників, а потім прикметники із сполучником, наприклад:

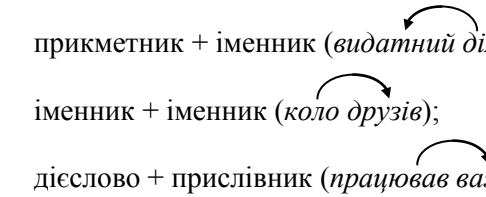
(порядні **банкіри** [ІС/ПЯ]), достойні **банкіри** [ІС/ПЯ], **порядні і достойні** [СУ]);

- якщо між ними безсполучниковий зв'язок, то сурядні зв'язки встановлюються з кожним із ад'єктивів та іменником (у будь-якій формі), а потім між самими ад'єктивами, наприклад:

порядні, достойні банкіри

(порядні **банкіри** [ІС/ПЯ]), достойні **банкіри** [ІС/ПЯ], **порядні**, **достойні** [СУ]).

Кожний тип словосполучення відображається у певному виді моделі. Модель словосполучення – це двоелементна формула, що відбиває один із типів зв'язку аналізованого слова з певним повнозначним словом, наприклад:



У тих випадках, коли прийменник (або сполучник) служить лише засобом зв'язку між двома повнозначними словами, він не вважається самостійним членом моделі. Таким чином, модель "дієслово + прийменник + іменник" (*працювати в уряді*) залишається двочленною, хоча складається з трьох слів. У проектованому словнику подаються чотири типи моделей: ядрові; неядрові (ад'юнктні, які відображають підрядні зв'язки); сурядні; предикативні.

АСА здійснюється за правилами – моделями, представленими у таблиці SyntaxRules у програмному середовищі Access, яка виконує роль диспетчера. Кожній моделі згідно з таблицею автоматично приписується певний код. Зокрема, за цією таблицею здійснюється "збирання" в один вузол складених морфологічних та синтаксичних явищ, наприклад: ГБ – аналітичний майбутній час (*буду читати*); ГЗ – аналітичний наказовий спосіб (*хай читає*); ГЧ – умовний спосіб (*читав би*); СЧ – складений числівник (*сорок три*); ПМ – складений підмет (*один з них*); ПС – складений присудок (*почав працювати*). Простим присудкам і безособовим формам дієслова приписуються коди: ПР та ГЧ відповідно. Причому в першому випадку далі ведеться пошук підмета, а в другому пошук продовжується за таблицею дієслівної валентності.

Типи синтаксичних словосполучень кодуються за частиномовною належністю: ІС – іменникове; АС – прикметникове; ДС – дієслівне; ЧС – числівникове; РС – прислівникове; ЗС – займенникове. За цією ж таблицею кодуються види синтаксичних зв'язків (vpr): КЗ – координація; ПЯ – підрядний ядровий; ПА – підрядний ад'юнктний; СУ – сурядний. Напрямки перевірки (праворуч / ліворуч) строго регламентуються набором правил для конкретного рядка, якими визначається і пріоритет у роботі групи правил.

Як свідчить тестування результатів роботи АСА, автоматично виділялися такі словосполучення і такі типи зв'язків, які повністю відповідають інтуїтивним уявленням носіїв української мови та експертів-лінгвістів. Отже, синтаксичну структуру речення автомат "зрозумів" правильно. І це відкриває широкі перспективи, зокрема можливість укладання частотного словника сполучуваностей української мови та здійснення автоматичного синтаксичного аналізу цілого речення. У свою чергу, правильний синтаксичний аналіз є запорукою створення автоматичного семантичного аналізу тексту.

1. *Вихованець І. Р.* Граматика української мови. Синтаксис / І. Р. Вихованець. – К., 1993. 2. *Дарчук Н. П.* ЭВМ в синтаксических исследованиях / Н. П. Дарчук // Использование ЭВМ в лингвистических исследованиях. – Киев, 1990. – С. 113–129. 3. *Дарчук Н. П.* Сегментация сложного предложения на составляющие предикативные части / Н. П. Дарчук // Синтаксический анализ научного текста на ЭВМ. – Киев, 1999. – С. 40–127. 4. *Загнітко А. П.* Основи українського теоретичного синтаксису. Частина 1 / А. П. Загнітко. – Горлівка, 2004. 5. Сучасна українська літературна мова. Морфологія. Синаксис. – К., 2010. 6. Русская граматика: в 2 т. Т. 2. Синтаксис. – М., 1980. – С. 21.

Рассматривается автоматическое представление синтаксической структуры предложения на уровне словосочетания: автоматическое выделение словосочетания, приписывание ему типа синтаксической связи (подчинительной, сочинительной, предикативной).

Ключевые слова: автоматический синтаксический анализ, словосочетание, синтаксическая связь, подчинительная связь, сочинительная связь, предикативная связь, ядерная подчинительная связь, адъюнктная подчинительная связь.

The article provides automated representation of syntactic analysis of sentence on the level of word-combination: automated marking of word-combination, characterising with a kind of syntactic relations (subordinating, coordinating conjunction, predicative relation).

Keywords: automated syntactic analysis, word-combination, syntactic relation, subordinating conjunction, coordinating conjunction and predicative relation.

Стаття надійшла до редакції 10.09.2012