1920s, and the period of the formation of theoretical foundations of lingvopersonology, which lasted until the 1980s. Although the term *linguistic person* itself appeared in the second of the allocated periods, scientists have not given them a complete interpretation.

Discussion: In the long run, the multi-vector path, to which the lingvopersonological studios at the end of the XX and XXI centuries dispersed, should be analyzed.

Keywords: lingvopersonology, personology, personalism, linguistic personality, anthropocentric paradigm.

Vitae

Illya G. Danyliuk, Candidate of Philology, Doctoral candidate and Associate Professor at Department of General and Applied Linguistics and Slavonic Philology in Donetsk National University named after Vasyl' Stus. His research areas include applied linguistics, natural language processing, corpus linguistics, and machine grammar.

Correspondence: i.danyluk@donnu.edu.ua

Наталія Дарчук

DOI 10.31558/1815-3070.2018.36.23
УДК 81`32

ДО ПИТАННЯ ПРО СТВОРЕННЯ АВТОМАТИЧНОГО СЛОВНИКА СЛОВОСПОЛУЧЕНЬ УКРАЇНСЬКОЇ МОВИ

У статті розглянуто засади створення автоматичного словника словосполучень української мови, створюваного на кафедрі української мови та в лабораторії комп'ютерної лінгвістики Інституту філології Київського національного університету імені Тараса Шевченка. Подано типологію синтаксичних зв'язків та семантичних відношень, на яких ґрунтовано створювану систему автоматичного виокремлення словосполучень, принципи семантичного розмічення словника. Описано експеримент з вилучення з корпусу публіцистичних текстів словосполучень з орудним відмінком, прокоментовано його результати.

Ключові слова: автоматичний семантичний аналіз, автоматичний синтаксичний аналіз, автоматичний словник словосполучень, комп'ютерна граматика, орудний відмінок, семантичне відношення, синтаксичний зв'язок.

Ідея взаємодії лексики і граматики, пов'язана з іменами В. Виноградова, Ю. Апресяна, Г. Золотової, І. Вихованця, К. Городенської, А. Загнітка, Р. Мразека, Р. Якобсона та ін., набуває актуальності у зв'язку з недостатньою розробленістю таких семантико-синтаксичних проблем, як: граматична і лексична валентність слів, типова частиномовна 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 та ідіоматичності тощо, а й практичними, зокрема створенням систем автоматичного аналізу тексту. З іншого боку, увагу до проблем семантизації синтаксису стимулюють прикладні проблеми: автоматизація 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, машинний переклад тощо.

Метою даного проекту є створення системи автоматичного виокремлення словосполучень з автоматичним приписуванням їм типу синтаксичного зв'язку і синтаксичного відношення та укладання частотного словника словосполучень із зазначенням цих типів інформації. Пропонований проект створюється на кафедрі української мови та в лабораторії комп'ютерної лінгвістики Інституту філології Київського національного університету імені Тараса Шевченка. Теоретичним підґрунтям є комп'ютерна граматика АГАТ із вбудованими процесорами опрацювання українськомовного тексту.

Зазначимо, що автоматичні процесори для вирішення текстових завдань за визначенням можуть бути тільки інтегральними комп'ютерними моделями мовної системи (так само як і в мові, де онтологічна система працює у взаємодоповняльному режимі до гносеологічного аспекту), її модель – діючий автомат аналізу тексту – ієрархічно вирішує всі необхідні завдання аналогічно до людини-лінгвіста, але робить це за правилами комп'ютерної граматики, яка складається за об'єктами опису з двох розділів – морфології і синтаксису та завершується звертанням до семантики, уникнути яку не можливо. Отже, зв'язок синтаксису із семантикою при вирішенні прикладних завдань логічний і необхідний.

Синтаксичний аналіз у системі АГАТ зумовлений у теоретичному плані тим, що виокремлення словосполучення з реченнєвої структури на великих різностильових масивах текстів дає можливість дослідникам української мови більш точно й аргументовано встановити синтаксичну і семантичну ємність цієї одиниці, у прикладному плані – розроблюваний автоматичний синтаксичний модуль аналізу українського тексту єдиний в Україні лінгвістичний ресурс, налаштований на синтаксичне розмічування, аналога якому немає, оскільки воно здійснюється тільки автоматично на базі повного автоматичного морфологічного аналізу зі знятою омонімією.

Ми надаємо великого значення словосполученню – міжрівневій лексико-морфолого-синтаксичій одиниці (Golovin 67). Кожний з компонентів словосполучення – **носій морфологічних ознак** із релятивною функцією, тобто властивістю приєднувати до себе словоформи і приєднуватися до них. Словосполучення – **номінативна лексична одиниця**, тому що поєднуються словоформи у словосполученні на основі сумісності лексичного значення кожного з них. Словосполучення, безперечно, – **синтаксична одиниця**, на основі якої формуються члени речення і синтаксичні категорії та явища. За всіма зазначеними аспектами словосполучення є важливою одиницею при породженні мовлення. На відміну від слова словосполучення реалізує відношення між елементами реальної дійсності на основі синтаксичних відношень між його членами. На відміну від речення, воно не є закінченим повідомленням – воно позбавлене часово-модального плану. Але словосполучення відіграє конструктивну роль при побудові речення, оскільки людина з дитинства запам'ятовує, а потім відтворює ті словосполучення, які були до неї «створені» відповідно до потреб оформлення думки. Цікавим є і те, що набір словосполучень демонструє риси індивідуальності, що мотивується соціальним статусом людини, її професією, освітою тощо. Психолінгвістичні дослідження показали, що реципієнт сприймає мовленнєвий потік як окремі смислові комплекси, адекватні словосполученням. Ці ідеї знайшли підтвердження і в квантитативній лінгвістиці – елементами словника мовної системи можна вважати такі одиниці, які зберігаються і відтворюються більшістю носіїв мови у цілісному вигляді (це слова, словосполучення) (Marchuk 92). Будь-яке прикладне лінгвістичне застосування (навчання мові, автоматичний аналіз і синтез текстів, машинний переклад тощо) потребує збереження всієї інформації щодо кожної лексеми мови з кількох причин: 1) треба знати, як різні обмеження, які накладаються мовою на сполучуваність, впливають на його вид і структуру; 2) спираючись на лексичне значення сполучуваних слів, спробувати зняти лексичну багатозначність або омонімію і «навчити» цьому комп'ютер. Отже, «навчанню» комп'ютера передуватиме автоматизований аналіз визначення синтаксичної інформації.

Словосполучення, виділені в корпусі текстів, повинні зберігатися при кожному окремому слові в електронному словнику поряд з іншою інформацією про слово. Класи словосполучень мають бути інвентаризовані з точки зору того, які типи словосполучень можливо і необхідно вносити до словника (Bolshakov 358). Оцінивши ці ідеї, ми дійшли висновку про необхідність їх реалізації в комп'ютерній граматиці. Це спонукало нас до розроблення лінгвістичного і програмного забезпечення частотного словника словосполучень, який у подальшому можна буде використовувати для семантичного аналізу тексту, а також як готовий продукт для лінгвістичних досліджень.

Лінгвіст потребує систематизації матеріалу у вигляді словників, зокрема словника словосполучень, щоб мати змогу докладно вивчити синтаксичні конструкції та їх залежність від лексико-семантичних характеристик їх компонентів і дати відповідь на питання: як відбувається ця взаємодія, які одиниці беруть у ній участь? Реєстр вилучених з Корпусу української мови словосполучень, з одного боку, є джерелом для наступних лінгвістичних досліджень, з іншого, – використавши семантичний потенціал у вигляді функціонально-семантичного аналізу, проілюструвати устрій українського синтаксису, можливо, визначити нові дискусійні моменти, які потребують обговорення.

Це електронний словник словосполучень української мови, в якому задіяні: 1) морфологічні ознаки (частиномовні та категорійні); 2) синтаксичні ознаки (зв'язки та відношення); 3) лексичні ознаки у вигляді таксономічних семантичних класів лексико-семантичних варіантів слів, які утворюють словосполучення (http://www.mova.info). Саме в систематизації словосполучень за кожною з цих ознак, ілюстрацією в реченні синтаксичних та семантичних властивостей ми вбачаємо занурення у глибинний синтаксис української мови. Такий проект створюється вперше в україністиці і на питання доцільності опису синтаксичного устрою у такому вигляді можна відповісти так: наукова доцільність вимірюється тим, як такий словник може бути використаний в подальших наукових розвідках, а також у прикладних завданнях – автоматичному реферуванні, машинному перекладі, автоматичному індексуванні, машинному навчанні. Щодо останнього – уточнюємо: експеримент з укладання словника здійснювався автоматизовано за допомогою спеціально укладених програм, які на підставі баз даних, що взаємодіють між собою, дають змогу лінгвістові працювати з кожним словосполученням і реченням, в якому воно зустрілося. Результати накопичуються, а потім після відповідної корекції програми запускаються на нових текстах, і машина автоматично тегує кожне словосполучення, приписуючи йому відповідну семантичну інформацію. Вичитування відбувається доти, доки машина не «навчиться» здійснювати семантико-синтаксичний аналіз словосполучення самостійно. Безперечно, система інформації, закладена в такого роду дослідження, є обговорюваною в сучасній науці і ще не до кінця вирішеною, тому, з гносеологічної точки зору, цікавим та корисним є досвід й отримані результати.

Логіка такого підходу відображає певну сходинку, в широкому смислі – в пізнанні будови матерії, а у вузькому – полягає в тому, що синтаксичний устрій мови влаштований за аналогією до навколишнього фізичного світу: складається з елементарних частинок, які організуються різноманітними, але регулярними комбінаціями елементарних одиниць, далі неподільних, але на синтаксичному рівні. І від осмислення того, що собою являють елементарні синтаксичні одиниці, можна зрозуміти більш складні побудови – речення, а відтак – текст як ціле. Отже, встановлення елементарних одиниць – частин більш складних явищ, з'ясування системних зв'язків між ними, а також між елементарними одиницями і складним цілим постали зараз в лінгвістичній науці, щоб відтворити наукову теоретичну картину українського синтаксису.

На «вході» системи автоматичного аналізу тексту кожне речення цілісного тексту обробляється автоматично трьома процесорами:

- морфологічним (кожній словоформі присвоюється частиномовний код і категоріальні граматичні морфологічні ознаки; граматична і лексико-граматична омонімія знімається автоматично);
- синтаксичним (виокремлюється словосполучення, йому присвоюється тип синтаксичного зв'язку, тип синтаксичного відношення на даному етапі дослідження автоматизовано);
- семантичним (кожному члену словосполучення автоматично приписується код семантичного таксону і семантичне відношення) (автоматизовано на цьому етапі дослідження).

На «виході» одержуємо речення, з якого вичленовується словосполучення, кожному з них приписується кортеж трьох інформацій.

Зупинимося докладніше на другому і третьому пунктах дослідження.

Дамо таке робоче визначення, взяте за основу у нашому дослідженні: **словосполучення – це смислове та граматичне поєднання двох або більшої кількості слів на основі підрядного, сурядного або предикативного зв'язку** (Zahnitko 104). Вважаємо, що словосполучення – відносно самостійна одиниця мови, виокремлювана у межах речення, яка створюється за правилами поєднання слів, виявляючи при цьому валентні властивості головного слова, реалізується у мовних моделях, які відтворюються у мовленні. Словосполученнями не вважаються складені аналітичні поєднання слів: 1) сполуки іменника з прийменником (*в Парламенті*); 2) складені (аналітичні) форми слів (*буду читати = читатиму, більш дорослий*), оскільки віднесення аналітичних сполук до словосполучень є некоректним, адже словосполучення витворюється поєднанням двох або більше повнозначних слів і виступає номінативною одиницею; 3) фразеологізми (*бити байдики*).

Завданням АГАТ-синтаксису є виявлення всіх різновидів сполучуваності – **предикативної, підрядної і сурядної** – кожного слова у тексті. Граматичні характеристики словосполучення безпосередньо залежать від того, до якої частини мови належить його ключове слово. Лексико-граматична природа слова визначає його здатність сполучатися з іншими словами. Відповідно до цього словосполучення поділяють на **іменникові, прикметникові, займенникові, числівникові, дієслівні та прислівникові**. За нашою концепцією АСА, виокремлення словосполучень базується на граматиці валентностей, а саме підграматиці для дієслова (**31 206 правил**), підграматиці для іменника (**40 023**), підграматиці для ад'єктива (**6 205**), а також на словнику фразеологізмів (близько **27200** одиниць). Комп'ютерні підграматики валентностей зазначених частин мови побудовані нами за єдиним принципом: вказано лексему, прийменник, який бере участь у керуванні, та відмінок іменникової словоформи у вигляді двобуквеного коду, прийнятого в АГАТ-граматиці (Darchuk).

Сполучення мовних одиниць (напр., дієслів з іменниками, іменників з іменниками тощо) можуть належати до одного з трьох основних типів залежно від того, наскільки жорстко пов'язані між собою елементи словосполучення. Сполучення, де ключове слово потенційно може поєднуватися з будь-яким словом тієї або іншої частини мови, належить до типу вільних сполук. Напр., дієслово *читати* потенційно сполучається з багатьма іменниками – *читати* можна *книгу, вірш, фразу, доповідь, пресу, лекцію, газету, оголошення, літературу* в тому числі з абстрактними: *читати думки, настрої.* На іншому полюсі знаходяться фразеологізовані сполуки – наприклад, *байдики* сполучається тільки з дієсловом *бити.* Між цими полюсами знаходиться тип так званих колокацій, невільних сполучень, які до ідіом (фразеологізмів) не відносяться. До інвентарних одиниць електронного словника фразеологізмів відносимо як безумовні фразеологізми (*бити байдики*), так і колокації – фразеологічні звороти, значення яких з погляду сучасної мови зовсім не мотивується, не виводиться із суми значень слів-компонентів (*брати участь, діаметрально протилежний*); менш ідіоматичні (*глибоко ображений*), а також прислів'я (*хочеш миру – готуйся до війни*). Не будемо зупинятися на лінгвістичних проблемах фразеологізації уведених до словника одиниць, але їх всіх об'єднує те, що вони, за Р. Рогожніковою, виступають під маркою «сполучень, еквівалентних слову» (Rogozhnikova). Щодо дискурсивних слів (сполучень), а також складених службових слів – вони підключаються до відповідних списків ще на початку СА.

За складом словосполучення поділяються на **прості, складні** та **комбіновані** (Zahnitko 104). За формально-синтаксичними й семантико-синтаксичними ознаками І. Вихованець поділяє словосполучення на прості й ускладнені (Vykhovanets' 195). Просте словосполучення складається з двох повнозначних слів й утворюється на основі одного синтаксичного зв'язку (*яскраве світло, кінець місяця*). Прості словосполучення є основою для утворення ускладнених словосполучень, які являють собою різноманітні об'єднання елементарних словосполучень на основі підрядного й сурядного зв'язків. Ускладнені словосполучення утворюються на основі трьох і більшої кількості підрядних зв'язків, зумовлених валентною сполучуваністю головного слова (*потиснути руку депутатові* тощо), а також у сурядних словосполученнях відкритої структури. Такі словосполучення з єднальними або розділовими відношеннями можуть мати у своєму складі з формальної точки зору необмежений ряд сурядних компонентів (Vykhovanets' 196). Ми розглядаємо тільки **прості бінарні словосполучення**, які можуть бути поширені в ускладнені, оскільки при визначенні їх складу потрібний аналіз смислової структури. Наприклад, немає ніяких формальних підстав виокремити поширені словосполучення ускладненого типу, в яких поєднуються опорні іменники із залежним прикметниковим узгодженням і неузгодженим іменниковим у формі родового відмінка: *напружена праця президента* тощо. В них поєднуються такі форми підрядного зв'язку, як узгодження і прилягання, узгодження і керування. За нашою методикою

аналізу буде виділено: *напружена праця* + *праця президента*. Оскільки у процесі аналізу зберігається номер кожної словоформи у базі даних (далі БД), можна подавати на екран «склеєне» з елементарних, простих словосполучень ускладнене. Натомість для ускладненого словосполучення, в якому поєднані опорний іменник із двома залежними прикметниками, передбачено два ступені членування: перше членування – виокремлення іменникового словосполучення із залежним прикметником; друге членування – перший прикметник підпорядковується не іменникові, а елементарному іменниковому словосполученню, напр., *тривала / важка праця*. Формальною ознакою для виділення такого ускладненого словосполучення є відсутність коми або єднального сполучника. Тоді одержимо словосполучення: *тривала важка праця* (стрілки від *праця*). При цьому між *тривала* і *важка* не визначається сурядний зв'язок.

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

Для кожного слова визначаються такі зв'язки: **підрядні, предикативні і сурядні,** оскільки вони відповідають відтворенню загальної системи відношень між компонентами описуваної ситуації у реченні. Віднесення до словосполучень тільки сполук із підрядним прислівним зв'язком, не є адекватним з точки зору складників речення (Vykhovanets' 50). Ми відмовилися і від традиційних видів підрядного зв'язку – узгодження, керування та прилягання, виходячи із широкого розуміння поняття синтаксичного зв'язку: це будь-які поєднання з головним словом відмінкової форми іменника. При узгодженні залежне слово уподібнюється головному в граматичних формах, а при приляганні воно, не маючи форм словозміни, приєднується до головного за змістом.

У комп'ютерній граматиці підрядні зв'язки поділяються на **ядрові** і **неядрові**. **Ядровим** називаємо такий зв'язок, при якому аналізоване слово є керувальним, головним. Напр., у реченні *Від економічної кризи сильно постраждали майже всі європейські держави* визначаємо 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 – це такий зв'язок, при якому аналізоване слово є залежним, керованим. У наведеному прикладі неядрові зв'язки спостерігаємо між словами *економічної* (залежить від *кризи*), *сильно* (залежить від *постраждали*), *європейські* (залежить від *держави*) тощо.

Предикативний – це зв'язок між основними компонентами речення «підмет – присудок», який грунтується на двобічній залежності головних членів речення, тобто на їх взаємозалежності. У наведеному реченні предикативний зв'язок спостерігаємо у сполученні *постраждали держави*. У таких випадках жодне зі слів не можна вважати домінувальним або підрядним, оскільки вони обидва є домінувальними. Якщо підмет і / або присудок є складеним, то визначається підрядний зв'язок для аналізованого слова, напр.:

Встановлюємо предикативний зв'язок між *двоє студентів* і *почали скандувати*. У сполуці ***двоє*** *студентів* ядровим буде форма ***двоє***, яка домінує над формою *студентів* (*студентів*, відповідно, має неядровий зв'язок); у складеному присудку ***почали*** *скандувати* ядровим елементом буде ***почали***, який домінує над *скандувати*. Те саме стосується іменного складеного присудка: *Він став студентом.* 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*.

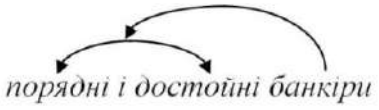
Сурядний – це зв'язок, при якому слова не є ані домінувальними, ані домінованими по відношенню одне до одного. Вважається, що два слова перебувають у сурядному зв'язку, якщо кожне з них підпорядковано одному і тому ж третьому слову, або якщо вони пов'язані між собою через сполучник, або відокремлені один від одного комою. Формальною підставою для виявлення сурядного ряду є перевірка кодів аналізованої пари слів за таблицею сурядності (див. Darchuk, додаток Б.9).

Автоматичне визначення сурядного зв'язку регламентується такими правилами:

1) Сурядний зв'язок установлюється між словами, а не між словом і зворотом або синтаксичною конструкцією. Напр., сурядний зв'язок наявний між словами *чесними* і *прозорими* (*Вибори були* ***чесними і прозорими***) і відсутній у такому випадку: *Президент був задоволений і в гуморі.*

2) Щодо зв'язку ад'єктивів (прикметників, дієприкметників, займенників), які виконують ту саму функцію, прийняті такі домовленості:

а) якщо між ними є сполучник, то напрямок зв'язку визначається: від головного слова до кожного з прикметників, а потім прикметників із сполучником, напр.:

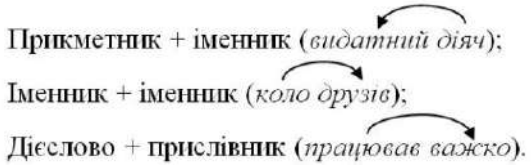


(порядні **банкіри** [мітка ІС/ПЯ], де ІС – іменникова сполука, ПЯ – ядрове), достойні **банкіри** [ІС/ПЯ], **порядні і достойні** [СУ], де СУ – сурядний зв'язок);

б) якщо між ними безсполучниковий зв'язок, то сурядні зв'язки встановлюються з кожним із них та іменником (у будь-якій формі), а потім між прикметниками, напр.: *порядні, достойні банкіри* (порядні **банкіри** [мітка ІС/ПЯ]), достойні **банкіри** [ІС/ПЯ], **порядні, достойні** [СУ]).

В усіх наведених трьох прикладах весь сурядний ряд видається на екран. Додамо також, що зручний інтерфейс, розроблений для користувача, дає можливість контролювати і коригувати результати АСА.

Кожний тип словосполучення відтворюється у певному типі моделі. Модель – це двоелементна формула, що виражає один із типів зв'язку аналізованого слова з певним повнозначним словом, напр.:



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, оскільки прийменник є формальним структурним елементом моделі.

При автоматичному укладанні словника словосполучень на будь-якому тексті Корпусу української мови одержуємо **чотири типи моделей**: **ядрові; неядрові** (ад'юнктні, які відображають підрядні зв'язки); **сурядні; предикативні.**

Типи словосполучень за морфологічним вираженням головного слова. Оскільки текст тегується морфологічним модулем, тобто встановлюються частиномовні і категорійні ознаки, а також виділяються синтаксичні словосполучення із зазначенням ядрового, головного компонента, то всі словосполучення поділяються на: 1) субстантивні (головне слово – іменник): *громадянське суспільство, депутат першого скликання*; 2) ад'єктивні (прикметник): *надзвичайно цікавий*; 3) нумеральні: *перший за списком*; 4) прономінальні (займенник): *хтось із студентів*; 5) дієслівні: *прагнути перемоги*; 6) адвербіальні (прислівник): *далеко від України*.

Щодо типів словосполучень за ступенем злитості їх компонентів, то вони поділяються на вільні (*будувати дім*) та фразеологічно зв'язані (*пекти раків*), а синтаксично зв'язані (*група крові*) поки що відносимо до вільних, оскільки їхня цілісність пізнавана тільки у межах речення, а експеримент з вилучення словосполучень здійснювався автоматично, причому сполуки були бінарними. Без втручання лінгвіста не обійтися – потрібна «ручна» робота.

Синтаксичні відношення у словосполученні розглядаються як смисловий зв'язок, який визначається характером відношення залежного слова до головного, тобто це смислові відношення, які характеризують синтаксичне значення словосполучення.

Типи словосполучень за характером смислових відношень між компонентами словосполучення поділяються на: 1) атрибутивні, 2) об'єктні, 3) суб'єктні, 4) обставинні, 5) комплетивні, 6) апозитивні.

Атрибутивні відношення. Атрибутивними словосполученнями виступають такі, в яких головне слово позначає предмет, а залежне окреслює його ознаку, якість предмета і відповідає у більшості випадків на запитання *який? чий?: виняткова причина, парламентська дискусія*. Автоматичне встановлення цього типу відношень не являє труднощів, оскільки на основі морфологічного аналізу модель A+N, де A – прикметник, займенник-прикметник, дієприкметник, які узгоджуються в категорійних ознаках, завжди буде кваліфікуватися як атрибутивне відношення.

Але атрибутивні відношення часто виникають і при сполученні іменника з іншими частинами мови:

а) іменник + іменник без прийменника: *криза уряду, наказ губернатора;*

б) іменник + іменник з прийменником: *будинок з дерева;*

в) іменник + інфінітив: *бажання вчитися*;

г) іменник + прислівник: *життя по-європейськи.*

Однак та сама модель може виражати різну семантику. Так, словосполучення, побудовані за схемою "ім. + ім. у р.в.", можуть реалізовувати атрибутивні й об'єктні відношення. У цих випадках на допомогу приходять семантичні теги, які присвоєні лексико-семантичним варіантам слів (ЛСВ) на семантичному етапі при створенні семантичного процесора (про це докладніше нижче). Кожне ЛСВ має семантичний код для кожної частини мови семантичної таксономії, тому перевіркою належності до певного таксону забезпечується можливість адекватного оцінювання типу відношення. Так, означенням залежний компонент у моделі "ім. + ім. у Р.в." виступає тоді, коли головне слово називає:

а) предмет, а залежне – особу, якій цей предмет належить: *будинок президента*;

б) частину предмета, а залежне – весь предмет: *сторінка контракту;*

в) сукупність предметів, а залежне – предмети, з яких ця сукупність складається: *країни світу;*

г) опредметнену дію, а залежне позначає діяча: *розпорядження міністра;*

ґ) ознаку, а залежне – особу або предмет, яким властива ця ознака: *цікавість виборців і под.*

Об'єктні відношення. Об'єктні відношення в словосполученні виникають при такій взаємодії компонентів, коли головне слово вимагає свого поширення об'єктом (на який спрямована чи з яким пов'язується дія): *затвердити закон, схвалений урядом* і под.

В об'єктних словосполученнях залежний компонент може позначати:

а) прямий об'єкт дії*: виголосити доповідь*;

б) знаряддя дії: *удар ногою*;

в) дійову особу: *підтриманий депутатами*;

г) співучасника дії: *розмовляти з колегами*;

ґ) об'єкт, що зумовлює зміст головного слова: *шкодувати за минулим*;

д) заклад або колектив, до яких особа має відношення: *завідувач кафедри.*

Об'єктні відношення виникають і при сполученні дієслова з інфінітивом: *наказати відступати, запросити виступити* та ін. Це, як правило, **дієслова руху або волевиявлення**.

Об'єктні відношення реалізуються і в словосполученнях, де головне слово виражене прикметником, а залежне іменником або іменниково-прийменниковим комплексом: *сильний духом, схильний до популізму* і под. Значення об'єкта залежне слово має в тих випадках, коли головне виступає віддієслівним іменником, який співвідносний з перехідним дієсловом, що вимагає зн.в.: *захист Вітчизни, виголошення доповіді // захищати Вітчизну, виголосити доповідь* та ін.

Суб'єктні відношення. Суб'єктні відношення наявні в словосполученнях, у яких залежна форма позначає чітко вираженого суб'єкта дії, діяча. Поки що прийнято гіпотезу: якщо залежний від дієслова іменник належить до семантичного таксона «ОСОБА», то автоматично приписується суб'єктне відношення, хоча можлива омонімія відношень: суб'єктних / об'єктних / атрибутивних. Такі випадки потребують додаткового дослідження, зокрема семантики дієслова – приналежності до певного таксономічного класу, адже суб'єктна семантика великою мірою залежить від лексико-граматичної природи головного слова, яким виступає дієслово: *приїхав депутат – приїзд депутата, ураган знищив – знищено ураганом*. Можливий збіг в одній формі двох видів відношень – суб'єктних та об'єктних, так звана синтаксична омонімія: *характеристика учня – учень характеризує* й *учня характеризують*.

Обставинні відношення. Обставинні відношення – це такі відношення, що характеризують чи кваліфікують дії (або ознаки), які за семантикою поділяються на відношення:

а) часові: *повернутися вчасно;*

б) просторові: *жити в місті;*

в) мети: *приїхати відпочити;*

г) причинові: *вчинити наперекір;*

ґ) наслідку: *зустрічатися на радість;*

д) означально-обставинні: *хоробро битися, гаряче говорити* і под.

Комплетивні (доповнювальні) відношення.

Комплетивні (доповнювальні) відношення виникають у синтаксично зв'язаних (цілісних) словосполученнях, коли головне слово вимагає смислового доповнення: *два кольори, вид спорту, ставати зеленим*. Словосполучення з цими відношеннями переважно виконують роль одного члена речення.

Апозитивні відношення. Апозитивні відношення охоплюють такі словосполучення, в яких обидва компоненти мають тотожну частиномовну приналежність і постпозитивний компонент характеризує головний за певною ознакою: *воїн-ветеран* та ін.

Значній кількості словосполучень притаманна синкретичність семантико-синтаксичних відношень, тобто в межах певної форми поєднується декілька значень. Так, можуть поєднуватися:

- атрибутивні й об'єктні відношення: *миска з пиріжками, дума про Україну, написання книжки*;
- атрибутивні й обставинні: *мандрівка в гори, повернення з міста, позиції обабіч шосе;*
- об'єктні й обставинні: *припухлий від сну*.

Однак шляхів зняття омонімії синтаксичних відношень ми поки що не встановили, тому в таких випадках можна приписувати коди омонімічних відношень або при накопиченні таких прикладів звернутися до семантики членів сполуки.

Зупинимося на семантичному блоці АГАТ-граматики, який являє собою семантичну розмітку словника – бази майбутнього семантичного аналізу українськомовного тексту, який впроваджується поетапно. Нею буде розмічено всі лексеми частотних словників, наявних у Корпусі українськомовних текстів різних стилів: художнього, публіцистичного, ділового, крім наукового – для різних підмов створюються тезауруси із системою відношень між термінами кожної терміносистеми. Наразі опрацьовано частотний словник публіцистичних текстів, генеральна сукупність яких у Корпусі перевищує 17 млн. слововживань. Укладено частотний словник цих текстів, обсягом 40 тис. різних лексем, лексику якого розподілено за частинами мови: словник іменників, словник дієслів, словник ад'єктивів, словник прислівників. Для кожної частини мови створюється своя семантична розмітка. Одиницею семантичного опису є не множина слів, а поняття, які відображають класи суспільнозначущих сутностей, розрізнюваних людьми, а **лексеми** у словнику відіграють роль **вербалізаторів понять.**

За основу було взято таксономію Національного корпусу російської мови як апробовану вже на корпусі текстів російської мови (Kustova; Krasil'shhik), яка при роботі з класифікацією лексем української мови зазнала

деяких змін. Лінгвістична таксономія – сукупність принципів і правил класифікації об'єктів, а також сама класифікація. Таксономія передбачає систематизацію як онтологічний результат, що частково відображає ієрархічну організацію, а для кожної частини мови розроблено свою таксономію зі своїм набором таксонів.

В кожній поняттєвій групі слова спочатку розташовуються за частинами мови (іменники, дієслова прикметники, прислівники тощо), а в межах одержаних граматично однорідних серій слів – за смисловою близькістю, що ілюструє лексичне багатство живої мови у всьому обсязі як єдине ціле. Таким же єдиним є й реальний світ, який відображається у лексиці, незважаючи на свою багатоплановість: всі події, предмети, процеси можуть розглядатися як прояв рухомої матерії, а їх взаємозв'язок підкреслюється наявністю цілого ряду об'єктивних законів руху матерії, які мають універсальний характер. Корелятом єдиного світу в пізнанні є єдине знання, закріплене у поняттях. Однією з форм фіксації понять є лексика мови (Karaulov).

Таксономічна класифікація лексики всіх частин мови складається з трьох частин: синоптичної, аналогічної та алфавітної.

Синоптична частина являє собою поняттєву класифікацію змістовної сторони лексики української мови. Наприклад, весь смисловий континуум для іменників було поділено спочатку на три основні поняттєві класи: ВЛАСНІ ІМЕНА, ПРЕДМЕТНІ ІМЕНА І НЕПРЕДМЕТНІ ІМЕНА, які потім розпадаються на 97 поняттєвих груп: ВЛАСНІ НАЗВИ – 9 груп, ПРЕДМЕТНІ ІМЕНА – 11, НЕПРЕДМЕТНІ – 77. У структурі таксономії це виражається в ієрархії таксономічних категорій, пов'язаних відношенням послідовного включення від нижчого рангу до вищого. Наприклад, до таксону ВЛАСНІ ІМЕНА як до більш загального класу включаються: ***димінутиви*** (*Саша, Сашко*), ***імена*** (*Олександр*), ***назви установ*** (*Азовсталь*), ***персонажі*** (*Білосніжка*), ***по батькові*** (*Іванович*), ***прізвища*** (*Іваненко*), ***топоніми*** (*Київ, Оболонь, Сула*), ***торгові марки*** (*Шанель*).

До таксону ПРЕДМЕТНІ ІМЕНА входять: ***артефакти*** (*архаїка, бовван, божок*); ***будинки, споруди*** (*автостанція, альтана, базиліка*); ***інструменти***, ***пристрої***, які конкретизуються таким вкладенням: ***зброя*** (*шабля, пістолет*), ***інструменти*** (*молоток, голка*), ***меблі*** (*стіл, диван, шафа*), ***музичні інструменти*** (*піаніно, скрипка, бандура*), ***одяг, взуття*** (*капелюх, чоботи, плаття*), ***посуд*** (*чашка, виделка*), ***транспортні засоби*** (*автобус, сани, потяг*); ***їжа, напої*** (*ацидофілін, бабка, балабуха*); ***нагорода, відзнака*** (*клейнод, корона, кубок*); назви ***осіб*** *(абітурієнт, вчитель, офортист*) з конкретизацією – ***етноніми*** *(абхаз, вірменин)*, ***імена родинності*** (*бабуся, братик*) та ***надприродних істот***, (*бабай, берегиня, берендей*); ***простір і місце*** (*акваторія, анклав, арборетум*); ***речовини і матеріали*** (*агат, агломерат, адреналін*); ***рослини*** (*азалія, айва*); ***тварини*** (*акула, альбатрос, амеба*); ***тексти*** (*абетка, абонемент, авізо*) тощо.

До таксону НЕПРЕДМЕТНІ ІМЕНА входять назви, пов'язані з процесами абстракції, без якого неможливий самий процес пізнання, тобто розкриття сутності, виокремлення істотних сторін об'єкта пізнання, проникнення в глибину дослідження предмета, адже абстракція – результат гносеологічної діяльності людини. Результатом процесу абстрагування є фундаментальні поняття: **Буття; Відношення; Кількість; Якість; Рух; Час** тощо.

Таксон НЕПРЕДМЕТНІ ІМЕНА є найчисельнішим, до нього входять 76 поняттєвих груп: ***буттєва сфера; взаємодія і взаємовідносини*** *(взаємодопомога, ворожнеча, сутичка);* ***вигляд*** *(обличчя, досконалість, імідж);* ***влада*** *(всевладдя, автократія, ареопаг);* ***властивість людини*** *(порядність, безвілля, дотепність);* ***властивість предметів*** *(гравітація);* ***враження*** *(ефект, інтерес);* ***галузі, промислові, заклади, виробничі об'єднання*** *(авіабудування, автопідприємство, автоцентр);* ***гра*** *(піжмурки, покер, доміно);* ***дія, діяльність*** *(акція, авторизація, агробізнес);* ***енергія*** *(випромінювання, гідроенергія, електроенергія);* ***заклад, організація, установа*** *(автоінспекція);* ***заняття*** *(акторство, автосправа, бавовництво);* ***запах*** *(аромат, перегар, повів);* ***захід*** *(аукціон, вернісаж, вибори);* ***звання, посада*** *(лауреатство, канцлер, капрал);* ***звук, шум*** *(аплодисменти, передзвін, шум),* ***зміна стану або ознаки*** *(укріплення, затвердіння, осушення);* ***інтенсивність*** *(приплив, злива, зрілість);* ***кількість*** *(видаток, виробіток, густина);* ***колір*** *(фарбування, колорит, жовтизна);* ***контакт й опора*** *(дотик, обійми, адгезія);* ***культура і мистецтво*** *(бароко, античність, готика);* ***межа*** *(поверхня, кінець, кордон);* ***ментальна сфера*** *(знання, абстракція, уява);* ***метод*** *(ангіографія, анімація, антропометрія);* ***місцезнаходження*** *(місцерозташування, адміралтейство);* ***мовлення*** *(дискусія, поголос, репліка);* ***музика*** *(арія, арієта, біг-бенд);* ***наука*** *(агробіологія, аеродинаміка, лінгвістика);* ***норма, правило*** *(вимога, вада, виняток);* ***обов'язок*** *(повинність, наряд, правочинність);* ***одиниця виміру*** *(бал, кілограм, метр);* ***ознака*** *(антинародність, артизм, атрибут);* ***параметр*** *(висота, вантажопідйомність);* ***поведінка і вчинки людини*** *(нехлюйство, підлабузництво, непокора);* ***погляд*** *(анахронізм, антагонізм, антропоцентризм);* ***подібність*** *(відповідність, асиміляція, еквівалент);* ***подія, пригода*** *(авантюра, вихор, гроза);****політичний напрям, течія*** *(лібералізм, радикалізм, шовінізм);* ***поняття*** *(анахронізм, аномалія, ідея);* ***порядок*** *(гармидер, бедлам, дезорганізація);* ***посесивна сфера*** *(володіння, придбання, втрата);* ***почуття*** *(любов, ненависть, симпатія);* ***право, закон*** *(дискримінація, засудження, зречення);* ***природне явище*** *(блискавиця, завірюха, спека);* ***процес*** *(вагітність, газифікація, газоочищення);* ***психічна сфера*** *(абулія, азарт, безнадія);* ***результат*** *(ефект, зрозуміння, капітуляція);* ***релігія*** *(баптизм, відправа, бузувірство);* ***риси характеру людини*** *(бережливість, іронічність, ідейність);* ***розмір*** *(гігантоманія, дольник, калібр);* ***розташування об'єкта*** *(розміщення, розстановка, намітка);* ***рух*** *(біганина, винесення, хитання);* ***світло*** *(промінь, напівтемрява, ілюмінація);* ***середовище, оточення*** *(антураж, твань, інтим);* ***спорт*** *(спартакіада, акробатика, баскетбол);* ***сприймання*** *(слух, видимість);* ***стан, становище*** *(анархія, байдужість, безвихіддя);* ***суспільний устрій*** *(лад, кріпаччина,*

народовладдя); ***температура*** *(прохолода, холоднеча, нагрів);* ***тон, колорит*** *(барва, індиго, барвистість);****умова*** *(гарантія, дно, змога);* ***факт*** *(деталь, компромат, каскад);* ***фізичний вплив*** *(удар, обмолот);* ***фізичні риси людини*** *(врода, дряхлість, зріст);* ***фізіологічна сфера*** *(спрага, крововилив, судома);* ***фінанси, гроші*** *(долар, акциз, аліменти);* ***форма*** *(геометрія, кільце, конус);* ***характеристика*** *(ярлик);* ***хвороба*** *(ангіна, діабет, арахноїдит);* ***час*** *(весна, роковини, хвилина);* ***чинник, причина*** *(агент, імпульс, збудник);* ***явище*** *(анахронізм, архаїзм, антагонізм);* ***якість*** *(ваба, довершеність, енергоємність)* тощо.

Нам видається така класифікація поняттєвого змісту іменників природною та аргументованою. Ця природність досягається тим, що логічна у своїй основі класифікація коригується живими асоціативними зв'язками слів. Так, якщо ми звернемося до поняттєвого класу БУТТЄВА СФЕРА, то можна одержати набір слів, який стосуватиметься ***існування*** (лексеми: *життя, буття* тощо); ***початку існування*** (*виникнення, народження, формування, творення* тощо); ***припинення існування*** (*смерть* тощо), які становлять вкладені класи.

В **аналогічній частині** таксономії іменника кожна з 97 поняттєвих груп наповнюється словами, причому поняттєві групи слідують в алфавітному порядку заголовкових слів-понять. **Алфавітна частина** є не що інше, як звичайний тлумачний словник сучасної української мови, в якому слова в кожному із своїх значень не тільки супроводжується вербальним визначенням, але й вказівкою на їх місце в таксономічній класифікації.

Таксономічна класифікація дієслів відрізняється від іменникової тим, що в ній основними одиницями аналізу є не окремі слова у певному значенні, а певні семантичні групи дієслів, що пояснюється специфікою денотативного значення цих частин мови: іменник позначає узагальнено предмет, а дієслово – тільки дію, яка відбувається з цим предметом, або її стан.

В основі організації лексичного матеріалу принцип ієрархічності спостерігаємо тільки для таксону **БУТТЄВА СФЕРА**, до якого входять: ***існування*** (*жити, відбуватися*); ***початок існування*** (*виникнути, народитися, сформувати*); ***припинення існування*** (*вмерти, убити, ліквідувати*). Умовно до буттєвої сфери можна віднести таксони: **ЯКІСНИЙ СТАН** *(біднішати, біліти, псуватися)*; **ЗМІНА СТАНУ АБО ОЗНАКИ** *(дорослішати, багатіти, розширитися)***,** які в рамках класифікації не перебувають в ієрархічних відношеннях. Решту таксонів можна охарактеризувати в такий спосіб: таксони, в яких згруповано дієслова дії та діяльності: **РУХ** *(бігти, смикатися, нести*); **ПЕРЕМІЩЕННЯ ОБ'ЄКТА** *(винести, пересунути)*; **ПОМІЩЕННЯ ОБ'ЄКТА** *(асфальтувати, одягти);* **ФІЗИЧНИЙ ВПЛИВ на об'єкт** *(бити, витирати, колоти);* **ТВОРЕННЯ, СТВОРЕННЯ ФІЗИЧНОГО ОБ'ЄКТА** *(базувати будувати, варити);* **ЗНИЩЕННЯ ОБ'КТА** *(висаджувати, зарізати, спалити);* **РОЗТАШУВАННЯ ОБ'ЄКТА** *(покласти, сховати*); **МІСЦЕЗНАХОДЖЕННЯ ОБ'ЄКТА** *(лежати, стояти);* **ПОЛОЖЕННЯ ТІЛА У ПРОСТОРІ** *(валятися, вивертати, сидіти);* **МЕНТАЛЬНА СФЕРА** *(вірити, здогадуватися, знати);* **МОВЛЕННЯ** *(говорити, каламбурити, радити)***; СОЦІАЛЬНА ДІЯЛЬНІСТЬ** *(втручатися, досягати, ховатися);* **СУСПІЛЬНО-ПОЛІТИЧНА ДІЯЛЬНІСТЬ** *(агітувати, займатися);* **ПОВЕДІНКА ЛЮДИНИ** *(вередувати, колобродити);* **ФІЗІОЛОГІЧНА СФЕРА** *(гикати, кашляти);* **ЗВУК** *(гудіти, шелестіти)*; **ЗАПАХ** *(духмяніти, пахнути);* **СПРИЙНЯТТЯ** *(дивитися, нюхати, слухати);* **СВІТЛО** *(гаснути, променитися).* Дієслова, які характеризують відношення: **ВІДНОШЕННЯ** *(берегти, відколюватися);* **МІЖСОСОБИСТІСНІ ВІДНОШЕННЯ** *(аплодувати, балувати, боготворити);* **СОЦІАЛЬНІ ВІДНОШЕННЯ** *(допомагати, індустріалізувати);* **ПОСЕСИВНА СФЕРА** *(мати, подарувати, набути);* **КОНТАКТ Й ОПОРА** *(торкатися, обнімати, спиратися)* тощо.

Було розмічено 31860 ЛСВ іменників, 47554 ЛСВ дієслів, 55959 ЛСВ прикметників і 12778 ЛСВ прислівників (http://www.mova.info).

Безперечно, розв'язуючи завдання із семантичного розмічування, можна одержати величезну кількість інформації про властивості слів і конструкцій, яка, з одного боку, буде корисною для уточнення номенклатури таксономії, а з іншого – дає матеріал для теоретичних висновків й узагальнень.

Таксономічну класифікацію можна розглядати в аспекті зв'язку із вивченням синтаксичної та семантичної сполучуваності, граматичної семантики, сталих синтаксичних конструкцій або конструкцій з двох чи трьох елементів за заданими морфологічними чи семантичними ознаками тощо. У поєднанні з Корпусом таксономічний словник суміщатиме в собі редукований ідеографічний словник і контекстний словник слововживань. Це значить, що, з одного боку, кожне слово у ньому подаватиметься в оточенні всіх семантично близьких йому слів, а з іншого – кожне слово супроводжуватиметься описом його сполучуваності. Такий словник, який можна назвати синтетичним, може бути новим типом семантичного словника. Слід наголосити на великому значенні тлумачного словника. При створенні автоматичного семантичного словника української мови дефініція буде використана двічі: з одного боку, як носій характеристики означального, як семантичний еквівалент слова, з іншого – як модель синтаксичної сполучуваності аналізованого слова.

Опишемо експеримент, здійснений з вилучення словосполучень з корпусу публіцистичних текстів. На рис. 1 представлено інтерфейс програми з автоматизованого розмічування словосполучень.

Рис. 1. Інтерфейс програми з автоматизованого розмічування синтаксичних сполук

Ми поставили завдання вилучити всі словосполучення з орудним відмінком з метою вироблення методики роботи. На цьому етапі роботи опрацьовувалися всі конструкції з орудним відмінком – прийменникові та безприйменникові, – в яких ад'юнктом був іменник, а ядровим решта повнозначних частин мови. Ми вважаємо, що вибір мовцем певної відмінкової словоформи із семи можливих відмінків, крім лексичного значення словоформи і характеру синтаксичних зв'язків, пов'язаний зі значенням відмінка, який виконує різні семантичні функції: інструмента дії (*бити ломом асфальт*), знаряддя дії (*дістатися машиною*), суб'єкта дії (*газета видається* (ким? чим?) *університетом*), об'єкта дії (*дівчина знизала* (чим?) *плечима)*, предикативної ознаки (*він був студентом*) та ін. Саме багатофункціональність орудного відмінка привернула нашу увагу, при цьому ми не розмежовували його первинні і вторинні семантичні функції.

На рис. 1 представлено інтерфейс програми, де є такі етапи роботи. Натиснувши на кнопку «Вибрати», обираємо текст, назва якого висвічується зліва у вікні. Натискаємо «Наступне», і подається іменник в орудному відмінку, зафарбований зеленим кольором, в оточенні контексту (5 у пре- і постпозиції до аналізованого іменника). Кожна словоформа супроводжується морфологічною інформацією (частиномовною та категорійною). Справа у боксах при кожній словоформі є кнопка «Вибрати», натиснувши на яку, одержуємо семантичну інформацію у вигляді назв таксонів з прихованим семантичним кодом. Якщо їх кілька, то за контекстом вибираємо потрібний таксон для усіх членів словосполуки (і прийменникової у тому числі). По завершенні семантичного опрацювання натискаємо кнопку «Зберегти». Після цього переходимо до перевірки синтаксичних зв'язків, які машина встановила автоматично (якщо потрібна редакція, то за викидним списком її коригуємо). Останніми встановлюємо семантичні відношення: придієслівні, приіменникові, приад'єктивні. Запам'ятовуємо і переходимо до наступного словосполучення.

Таким чином було опрацьовано близько 2000 словосполук з орудним відмінком ад'юнкта (993 придієслівні, 604 приіменникові, 184 приад'єктивні), в яких сполучаються морфологічна, лексична, синтаксична і семантична інформації із контекстами вживання. За цією ж схемою опрацьовуватимуться всі інші відмінки. Цей ілюстративний матеріал, з одного боку, складатиме основу електронного частотного словника словосполучень, а з іншого – ляже в основу автоматичного розмічення за семантичним відношеннями.

Ми вважаємо, що для дослідження сполучуваності треба провести переконливий аналіз корпусних даних, на підставі яких слід відібрати словосполучення, реально оцінити межі вільних сполук (на нашу думку, вільні сполуки все-таки мають певні межі, отже, є до деякої міри невільними). Ідіоми та фразеологізми виключалися з аналізу, однак деякі колокації включалися, якщо ад'юнктний іменник міг сполучатися з іншими дієсловами.

Планується укласти словник з кількома «входами»: перший вхід – пошук за ядровим словом; другий – пошук за ад'юнктним; третій – пошук за таксономією ядрового, четвертий – пошук за таксономією ад'юнктного, що дозволить вивчати сполучуваність конкретних лексем, як ядрових, так і ад'юнктних за належністю до певного таксону, напр., сполучуваність дієслів з непредметними іменниками. Такий матеріал може стати основою для формування іншого виду словника – за семантичними функціями ад'юнктного іменника.

References

Bolshakov I. A., Gelbukh A. F., and Galicia-Haro S. N. Electronic dictionaries: for both humans and computers. *Text, speech, and dialog : second International workshop TSD'99,* Plzen, Czech Republic, Sept. 13–17, 1999. Eds.: V. Matousek [et al.]. Berlin: Heidelberg Springer, 1999. 365–368. Print.

Darchuk, Nataliya. *Komp`yuterne anotuvannya ukrayins'koho tekstu: rezul'taty i perspektyvy (Computer Annotation of Ukrainian Text: Results and Prospects)*. Kyyiv: Osvita Ukrayiny, 2013. Print.

Golovin, Boris. *Vvedenye v yazykoznanye (Introduction to linguistics).* Moskva: Vyssh. shk., 1966. Print.

Karaulov, Jurij. *Lingvisticheskoe konstruirovanie i tezaurus literaturnogo jazyka (Linguistic Design and Literary Language Thesaurus).* Moskva: Nauka, 1981. Print.

Krasil'shhik I.S., and Rahilina E.V. Predmetnye imena v sisteme «Leksikograf» (Subject names in the system "Lexicographer"). *Nauchno-tehnicheskaja informacija (Scientific and technical information)*. Serija 2. 9 (1992): 24–31. Print.

Kustova, G.I., Ljashevskaja, O.N., Paducheva, E.V., and Rahilina, E.V. Semanticheskaja razmetka leksiki v nacional'nom korpuse russkogo jazyka: princhipy, problemy, perspektivy (Semantic markup of vocabulary in the national corpus of the Russian language: principles, problems, prospects). *Nacional'nyj korpus russkogo jazyka: 2003-2005 (National Corpus of the Russian language).* Moskva: Izd-vo «Indrik». 2005. 155-174. Print.

Marchuk, Jurij. *Komp'juternaja lingvistika (Computational Linguistics).* Moskva: ACT: Vostok-Zapad, 2007. Print.

Rogozhnikova, Roza. *Tolkovyj slovar' sochetanij, jekvivalentnyh slovu* (*Dictionary of combinations equivalent to the word)*. Moskva: AST: Astrel', 2003. Print.

Vykhovanets', Ivan. *Hramatyka ukrayins'koyi movy. Syntaksys (Grammar of the Ukrainian language. Syntax).* Kyyiv: Lybid', 1993. Print.

Zahnitko, Anatoliy. *Osnovy ukrayins'koho teoretychnoho syntaksys (Fundamentals of Ukrainian Theoretical Syntax).* Horlivka: HDPIIM. 2004. Ch. 1. Print.

Надійшла до редакції 23 листопада 2018 року.

TO THE ISSUE OF CREATION THE AUTOMATIC VOCABULARY OF WORD COMBINATIONS OF THE UKRAINIAN LANGUAGE

Natalia Darchuk
Department of Ukrainian Language and Applied Linguistic,
Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

Abstract

Background: Being created within the Ukrainian Language Corpus (http://www.mova.info) the system of automatic phrases distinguishing is based on syntactic and semantic divisions of the corpus. Theoretical background of the system set-up is the computer grammar AGTA (automatic grammar analysis of the text) with embedded processor of Ukrainian text processing.

Purpose: The purpose of the research is to set up the system of automatic phrases distinguishing with automated attaching the type of syntactic connection and syntactic relationship to them and frequency dictionary of phrases arrangement indicating these types of information.

Results: The experiment on all ablative case phrases (prepositional and non-prepositional) abstracting from publicistic texts subcorpus of the Ukrainian Language Corpus is conducted. About 2,000 collocations consisting ablative case of an adjunct are processed (993 adverbial, 604 ad-substantive, 184 ad-adjectival).

Ablative case in the Ukrainian language is multifunctional. The choice of a certain case word form a speaker makes from seven possible, except lexical meaning of a word form and the character of syntactic relations, is connected with the meaning of the case which fulfils different semantic functions: the tool of action (Ukr. *бити ломом асфальт (to break concrete with a scrap)*), the instrument of action (Ukr. *дістатися машиною (to get by car)*), a subject of action (Ukr. *газета видається університетом (a newspaper is published by the university)*), the object of action (Ukr. *дівчина знизала плечима (a girl shrugged her shoulders))*, predicate features (Ukr. *він був студентом (he was a student)*) and others.

Discussion: We are convinced that to investigate co-occurrence and estimate free combinations' bounds significant corpus data, on ground of which phrases should be selected, analysis must be conducted (we consider that free combinations still have certain verges, thus, they are constrained to some extent). It is planned to compile a dictionary with some «entries»: the first entry is the search by a core word; the second one is the search by an adjunctive; the third is the search by taxonomy of a nuclear; and the fourth is the search by taxonomy of adjunctive. This enables the study of the particular lexical items (both core and adjunctive) co-occurrence by their belonging to a certain taxon, e.g. a verb and a non-substantive noun co-occurrence. Such material can become a basis for compiling another kind of dictionary – by semantic functions of an adjunctive noun.

Keywords: automatic semantic analysis, automatic syntactic analysis, automatic phrases dictionary, computer grammar, ablative case, semantic relation, syntactic relation.

Vitae
Natalia Darchuk, Doctor of Philology, professor, professor of Department of Ukrainian Language and Applied Linguistic in Taras Shevchenko National University of Kyiv. Her areas of research interests include applied linguistics and computational linguistics.
Correspondence: nataliadarchuk@gmail.com

Ганна Ситар

DOI 10.31558/1815-3070.2018.36.24
УДК 81'373.7:81'32

СТАТИСТИЧНИЙ АНАЛІЗ ЦІЛІСНИХ СЛОВОСПОЛУЧЕНЬ: НА МАТЕРІАЛІ УКРАЇНСЬКОГО НАЦІОНАЛЬНОГО ЛІНГВІСТИЧНОГО КОРПУСУ[1]

Стаття продовжує цикл публікацій, присвячених статистичному аналізу стійких одиниць української мови. За даними Українського національного лінгвістичного корпусу визначено ступінь невипадковості поєднання словоформ дво-, три- і чотирикомпонентних цілісних словосполучень української мови шляхом обчислення показника асоціації mutual information (MI).

Для всіх обстежених цілісних одиниць властива невипадковість поєднання словоформ (результати MI перебувають у діапазоні від 8,64 до 44,63). У межах одного корпусу текстів величина MI залежить від таких чинників, як абсолютна частота конструкції, абсолютна частота її компонентів, кількість компонентів і тип цілісного словосполучення.

Ключові слова: показник асоціації, фразеологічна одиниця, mutual information, статистика, цілісне словосполучення, українська мова.

Постановка проблеми, актуальність дослідження. Сучасна лінгвістика визначає статистичні дослідження як виключно корпуснобазовані, тобто вчені вважають, що вірогідні статистичні результати можна одержати тільки на підставі аналізу репрезентативного корпусу текстів. Статистичний аналіз стійких одиниць різних типів, виконаний на корпусному матеріалі, є важливим завданням лінгвістичної статистики, здатним дати чіткі кількісні критерії зарахування мовних одиниць до класу стійких, що можуть бути використані для їх автоматичної ідентифікації в корпусі текстів. Процедуру такого аналізу на матеріалі синтаксичних фразеологізмів української мови викладено у працях (Sytar, "Statystychni Kryteriyi Analizu Syntaksychnykh Frazeolohizmiv"; Sytar, "Statystychnyi analiz frazeolohizovanykh rechen…"; Sytar, "Syntaksychni frazeolohizmy v rozrizi konstruktsiinoi hramatyky").

Обчислення показників (індексів) асоціації (англ. association measures, measures of association) як метод визначення випадковості / невипадковості поєднання певних одиниць може бути застосований для різних типів конструкцій. Пропонована стаття присвячена статистичному аналізу цілісних словосполучень.

У трактуванні словосполучення маємо опертям традиційний підхід, згідно з яким його визначають як непредикативну синтаксичну одиницю, «компонентами якої є слово та форма слова або кілька форм слів, з'єднаних між собою підрядним синтаксичним зв'язком» (Zahnitko, "Slovnyk suchasnoyi linhvistyky: ponyattya i terminy", 1040).

У розгалуженій класифікації словосполучень цілісні[2] одиниці посідають особливе місце й охоплюють кілька структурних і семантичних різновидів. Вони становлять один із трьох типів словосполучень, які виділяють за ступенем злиття компонентів. За цією ознакою з-поміж словосполучень Анатолій Загнітко розмежовує:

1) вільні словосполучення;

2) синтаксично зв'язані словосполучення;

3) фразеологічно зв'язані словосполучення (Zahnitko, "*Teoretychna hramatyka ukrayins'koyi movy: Syntaksys*", 63).

На думку мовознавця, визначальними ознаками синтаксично зв'язаних (або нечленованих, неподільних, цілісних) словосполучень є такі: виконання ролі одного члена речення, наявність структури і граматичного

[1] Дослідження виконано в межах наукового проекту «Об'єктивна і суб'єктивна мовносоціумна граматика: комунікативно-когнітивний та прагматико-лінгвокомп'ютерний виміри» (0118U003137) Донецького національного університету імені Василя Стуса.

[2] В україністиці, крім терміна «цілісні словосполучення», у межах різних підходів до кваліфікації цих одиниць та створення різних класифікацій дослідники використовують також терміни «неподільні словосполучення», «синтаксично неподільні словосполучення», «семантично неподільні словосполучення», «нерозкладні словосполучення», «нечленовані словосполучення», «синтаксично нечленовані словосполучення» і под. (Zahnitko, Balko, Maksymiuk, Lychuk та ін.).