2. Крылова И. П., Гордон Е. М. Грамматика современного английского языка: Учебник для ин-тов и фак. иностр. яз.— 5-е изд.—М.: Книжный дом «Университет», 2000.— 488 с.— На англ. яз.
3. Рыбаков Ф. И. и др. Автоматическое индексирование на естественном языке/ Рыбаков Ф. И., Руднев Е. А., Петухов В. А. — М.: Энергия, 1980. — 160 с.
4. Карпіловська Є. А. Вступ до комп'ютерної лінгвістики / Донецький національний університет. — Донецьк: ТОВ «Юго-Восток, Лтд», 203. — 184 с.
5. Качалова К. Н., Израилевич Е. Е. Практическая граматика английского языка: В 2-х т. К.: Методика, 2000 г. — 368 с.
6. Положення про навчально-наукові та кваліфікаційні роботи студентів Київського державного лінгвістичного університету / Укл.: Болдирєв Р. В., Соловей М. І., Сажко Л. А., Спіцин Є. С. – К.: Вид. центр КДЛУ, 1999. – 46с.
7. M. A. Ganshina, N. M. Vasilevskaya English Grammar: Ninth edition revised – M.: Higher school publishing house, 1964. – 548p.
8. Освітні лінгвістичні Інтернет-портали: www.webster.commnet.edu; www.ego4u.com; www.paulnoll.com; www.vionet.hit.bg; www.wordpower.ws; www.algebra.com; www.fortunecity.com; www.ielanguages.com

О. Бугаков*
Український мовно-інформаційний фонд НАН України (Київ)
УДК 81'322.5

ВСТАНОВЛЕННЯ ЗОН ПРИЙМЕННИКОВИХ ЗВ'ЯЗКІВ В УКРАЇНСЬКОМУ ТЕКСТІ (МОДУЛЬ АВТОМАТИЧНОГО СИНТАКСИЧНОГО АНАЛІЗУ)

The principles of construcnting the algorythm for identification of prepositional connections zones in the Ukrainian text (module of parsing) are given. The algorythm has been constructed on the material of linguistic database as a subcorpus of the Ukrainian National Linguistic Corpus according to the analysis of prepositional connections zones in syntactic structure of the Ukrainian sentence.

Прийняте в Українському мовно-інформаційному фонді (УМІФ) поетапне здійснення синтаксичної розмітки лінгвістичного корпусу текстів базується на принципі часткової розмітки з використанням методу фракційного синтаксичного аналізу. Суть останнього полягає у побудові синтаксичної структури речення з окремих фракцій – синтаксичних конструкцій певного типу, а саме: ланцюжків слів, об'єднаних предикативним зв'язком, іменних безприйменникових конструкцій, прикметниково-іменникових конструкцій, прийменникових конструкцій (зон прийменникових зв'язків), на аналіз яких, власне, і спрямоване наше дослідження. Шляхом об'єднання синтаксичних конструкцій одержуємо повне представлення синтаксичної структури речення. Звідси й назва – фракційний синтаксичний аналіз.

Зона прийменникових зв'язків (ЗПЗ) містить прийменник, головне слово (тобто слово, яке керує прийменниково-іменниковою синтаксемою) і залежне слово (тобто слово, яке підпорядковується головному за допомогою прийменника).

Метою дослідження є побудова алгоритму встановлення зон прийменникових зв'язків в українському тексті (як окремого модуля автоматичного синтаксичного аналізу (АСА)) з урахуванням якісних та кількісних характеристик організації прийменникових зв'язків в українському тексті.

Виведення алгоритмічних правил визначення ЗПЗ передбачає проведення аналізу на репрезентативному матеріалі з метою забезпечення належного рівня достовірності результатів. Таким матеріалом слугував Український національний лінгвістичний корпус (УНЛК) обсягом 36 млн. слововживань, створений в УМІФі.

Програмне забезпечення УНЛК, побудованого як інформаційна система, дозволяє створювати спеціалізовані субкорпуси, орієнтовані на розв'язання поставлених завдань. За допомогою спеціально розробленої в УМІФі програми зазначені субкорпуси переводяться у формати баз даних з певною структурою, орієнтованою на проведення конкретних досліджень синтаксичної структури текста. Лінгвістичні бази даних (ЛБД), які виконують функцію інструмента і матеріалу дослідження мовного явища, структуровані за таким принципом: текстовим сегментам (контекстам), що містять конкретну мовну одиницю (прийменник), ставляться у відповідність визначені наперед диференційні ознаки, за якими здійснюється аналіз. Структурування ЛБД за полями, що

відповідають множині параметрів аналізу діагностуючих контекстів, та організація доступу до цих полів дозволяють автоматично класифікувати матеріал за кожним із параметрів та будь-якою їх комбінацією.

Інформаційною базою для визначення граматичних, позиційних та семантичних ознак компонентів ЗПЗ слугувала створена лінгвістична база зон прийменникових зв'язків (ЛБЗПЗ), сформована на субкорпусі обсягом 6 млн. слововживань. Обсяг бази – 20.768 речень. База структурована за полями, що відповідають множині параметрів, попередньо визначених прогнозуючими текстовими ознаками для алгоритмічного встановлення ЗПЗ при АСА. При формуванні полів цієї ЛБД виходили з того, що для визначення алгоритмічних правил АСА суттєвими і формально встановлюваними з тексту типами лінгвістичної інформації є: 1) морфологічна інформація текстових словоформ; 2) інформація про сполучувальні властивості граматичних класів слів у межах певних синтаксичних конструкцій; 3) дані про пунктуаційні засоби, які використовуються при структуруванні речення; 4) відомості про позиційні умови реалізації певних типів синтаксичних зв'язків; 5) лексико-статистична та 6) лексико-семантична інформація [4; 5]. Зазначеним типам інформації відповідають поля ЛБЗПЗ: 1) "Контекст", 2) "Довжина ЗПЗ", 3) "Перша позиція прийменника", 4) "Постпозиція ГС", 5) "Контактність ГС", 6) "ГС", 7) "Граматичний код ГС", 8) „Семантичний клас ГС", 9) "Контактність ЗС", 10) "ЗС", 11) "Граматичний код ЗС", 12) „Семантичний клас ЗС", 13) „Відношення", 14) „Ремарки" [3].

Виведені на основі ЛБЗПЗ якісні та кількісні характеристики організації прийменникових зв'язків в українському тексті [2] лягли в основу алгоритму встановлення зон прийменникових зв'язків в українському тексті. Їх також передбачається використати при ідентифікації прийменникових конструкцій на етапі синтаксичної розмітки УНЛК.

Розробка лінгвістичних алгоритмічних процедур в АСА будується з урахуванням принципів [1]: 1) доцільності (наприклад, для прийменників з перевагою постпозитивного розташування ГС слід починати пошук ГС саме з цієї позиції); 2) найбільшого виграшу (спочатку розглядаються зв'язки з найбільшою діагностуючою силою: наприклад, перевага контактного ГС-дієслова зумовлює початок перевірки саме з цієї ситуації); 3) частоти вживання (в алгоритмі частота прийменників враховується при визначенні ієрархії правил); 4) економічності (поступове ускладнення процедури перевірок).

Через те, що встановлене залежне слово може допомогти у пошуку головного, алгоритм встановлення ЗПЗ у тексті повинен починатися з визначення ЗС. По-перше, для залежного слова характерні обмежені синтагматичні контакти з прийменником, оскільки воно чіткіше визначене з погляду своєї позиції (переважно постпозиція стосовно прийменника) та граматичних характеристик (число класів, які можуть виконувати роль ЗС, обмежене), а отже, немає потреби залучати семантичну інформацію. По-друге, і це головне, результати цього етапу можуть використовуватися на наступних, зокрема, при використанні семантичних характеристик ЗС для пошуку ГС. Крім того, граматичні ознаки залежного слова можуть передбачати напрямок пошуку ГС і встановлювати порядок перевірки лівого і правого оточення прийменника.

При визначенні ЗС враховувалися результати попереднього аналізу функціонування прийменників у тексті: між прийменником і ЗС ніколи не може знаходитися жоден із компонентів предикативної пари; прийменник і ЗС не можуть знаходитися в різних предикативних частинах (ПЧ) складного речення.

Визначення ЗС починається з перевірки правого контексту прийменника на наявність у ньому іменника, займенника-іменника або числівника, які виконують функцію ЗС у 95, 9% усіх зафіксованих випадків у ЛБД: ***Навколо<PB> них<OT>*** *уже<Z0> сьогодні<T0> формується<YO> культ<MA> .<e>*

У випадку наявності одного з цих класів передбачається перевірка на узгодженість його відмінка з інформацією про відмінок, яким може керувати конкретний прийменник.

При негативному результаті перевіряється правий контекст на наявність у ньому цифри, абревіатури або скорочення, які переважно розташовуються в контактній позиції до прийменника: *Під час<PB> виступу<MB> в<PF> альма-матер<FF> Клінтон<JA> наголосив<VR> на<PF> необхідності<FF> продовження<NB> роботи<FB>* ***над<PE> ПРО<2A>*** *.<e>*

Для прийменників, у яких попередньо визначена можливість препозитивного розташування ЗС (*вслід/услід, навздогін/наздогін, навперейми, назустріч/навстріч,*

наперекір, *напереріз*, *на зміну*, *на противагу*, *ради*, *заради*), після перевірки правого контексту додатково перевіряється лівий на наявність у ньому контактно розташованих іменника та займенника-іменника у родовому (для *ради*, *заради*) або давальному (для інших прийменників) відмінках:

Правда<FA> йде<UO> ***Правді<FC> вслід<PC>*** *.<e>*

Після завершення роботи алгоритму пошуку залежного слова в алгоритм пошуку головного слова ЗПЗ передається інформація про місце розташування прийменника і ЗС в реченні і про граматичні характеристики останнього.

Межами інтервалу пошуку ГС при перевірці препозитивного контексту є: слово в початковій позиції речення або ПЧ, початок звороту (дієприкметникового або дієприслівникового), член предикативної пари, сполучення цифри з дужкою, яка вживається при переліку пунктів, підрядний або сурядний сполучник з попередньою комою. При пошуку ГС межі включаються в число об'єктів, які перевіряються на роль ГС. Межею постпозитивного пошуку є: присудок (оскільки для аналізу подається текст з установленим предикативним центром), останнє слово речення або ПЧ, для вставних конструкцій у дужках – дужка. При визначенні меж передбачено пропуск вставних конструкцій у дужках при лівому і правому пошуку і відокремлених зворотів – при лівому.

Пошук ГС починається з перевірки позиції прийменника. Якщо він знаходиться на початку речення або предикативної частини (ПЧ), головним словом визнається присудок (переважно виражений дієсловом, присудковим словом чи прикметником):

Незважаючи_на<PD> *дві<HV> поразки<FK>, ми<OS> не<Z0>* ***втратили<VU>*** *шансів<MI> знову<T0> вийти<VA> у<PD> чвертьфінал<MD> .<e>* – ГС у ЗПЗ прийменника *незважаючи на* виражене дієсловом *втратили.*

У випадку розташування прийменника не в першій позиції при виборі напрямку пошуку ГС повинні враховуватися дані про співвідношення ліво- та правобічних зв'язків конкретних прийменників. Як показав попередній аналіз, обидва способи розташування властиві майже всім прийменникам (крім *у справах*, *вглиб*, *на засадах*, *осторонь*, *у напрямі*, *усередину*, для яких було зафіксовано виключно лівобічне розташування ГС), тому правила аналізу повинні бути розраховані як на лівий, так і на правий контексти. Для прийменників з перевагою лівобічних зв'язків правила пошуку ГС у правому контексті повинні працювати в останню чергу, а для перерахованих вище прийменників в алгоритмі повинно передбачатися звернення до правого контексту як резервного правила, оскільки принципова можливість постановки ГС праворуч від прийменника не виключається. Для прийменників з перевагою правобічних зв'язків (*згідно*, *незважаючи на*, *пріч*, *згідно з*, *на відміну від*, *попри*) пошук ГС в правому контексті передує роботі правил, які аналізують лівий контекст. У цьому випадку відбувається пошук присудка праворуч від прийменника, роль якого переважно виконують дієслово, присудкове слово, прикметник чи дієприкметник:

Примітно<T0>, що<SS> ***згідно_з<PE>*** *латвійським<AE> законом<ME> про<PD> приватизацію<FD>, у<PF> країні<FF> не<Z0>* ***існує<UO>*** *заборонених<AT> до<PB> приватизації<FB> об'єктів<MI> …<e>*

Для інших прийменників пошук ГС починається з аналізу лівого контексту. Як показав аналіз всіх ЗПЗ, при розташуванні дієслова, присудкового слова, прикметника чи дієприкметника в контактній позиції до прийменника в 99, 9% випадків вони виконують роль ГС.

При лівобічному розташуванні ГС важливою є інформація про перевагу у прийменників дієслівного чи іменного керування, яка була отримана внаслідок попередньо проведеного аналізу. Для прийменників, які віддають перевагу дієслівному керуванню (переважна більшість прийменників), за алгоритмом передбачається, насамперед, пошук дієслова або присудкового слова у дистантній позиції стосовно прийменника, але в межах інтервалу пошуку, оскільки, як показав матеріал, прийменники, які мають частку дієслівного керування більшу, ніж іменного, в тексті надають перевагу віддаленим зв'язкам з дієсловом (або його формами), ніж близьким зв'язкам з віддієслівними іменниками на відміну від прийменників, у яких в тексті переважає іменне керування: *Канівські<AS> судді<MH> дійшли<VU> висновку<MB>, що<SS> Володимир<JA> скоїв<VR> убивство<ND> у<PF> стані<MF> афекту<MB>, і<SC> звільнили<VU> його<OB> з-під<PB> варти<FB>, де<T0> він<OA>* ***перебував<UR>*** *під_час<PB> слідства<NB>* ***протягом<PB>*** *10<1A> місяців<MI> .<e>*

– головним словом у ЗПЗ прийменника *протягом* є не контактно розташований іменник з процесуальним значенням *слідство*, а дієслово *перебувати*, оскільки цей прийменник віддає перевагу дієслівному керуванню (84, 2% вживання в ЛБЗПЗ).

Для прийменників, у яких за попередніми даними частка іменної залежності більше, ніж дієслівної (*у справах*, *у галузі*, *з боку*, *на зразок*, *щодо*, *між*, *у справі* тощо), спочатку відбувається перевірка на пошук іменника з процесуальним значенням (розташованого як контактно, так і дистантно стосовно прийменника) у межах інтервалу пошуку:

Щоправда<T0>, в<PF> Головному<AR> столичному<AR> ***управлінні<NF> у_справах<PB>*** *захисту<MB> прав<NI> споживачів<MI> переконані<CS> : не<Z0> можна<X0> отоварюватися<YA> на<PF> стихійних<AX> ринках<MM> .<e>*

Виходячи з властивості проективності речень української мови (тобто неможливості перетину синтаксичних стрілок у структурі речення), виключається можливість іменного керування прийменником при вклинюванні між ними присудка, вираженого дієсловом, присудковим словом, прикметником чи дієприкметником. Відповідно, якщо ГС прийменника – іменник, то він може знаходитись тільки в інтервалі між присудком та прийменником. Якщо в ланцюжку слів ліворуч від прийменника не знайдено іменника, який задовольняє ознаки ГС-імені, а в інтервалі знайдено один з трьох класів, то останній і буде визначений як ГС прийменника.

Описаний вище принцип розташування правил є спільним для всіх прийменників. До спільних правил відноситься також використання семантико-граматичної інформації, зокрема, перевірка словоформ, розташованих ліворуч від прийменника, на входження до класів *істота* і *дія*. Перший формально може бути виведений за номером парадигматичного класу, що визначається для кожної текстової словоформи на етапі попереднього АМА тексту. У результаті проведеного аналізу було визначено прийменники, у яких функцію ГС може виконувати іменник-істота (*без*, *в/у*, *від*, *для*, *до*, *з*, *за*, *на*, *під*, *по*, *біля*, *з боку*, *над*, *поза*, *понад*, *протягом*, *у галузі*, *внаслідок*, *з погляду*, *з-за*, *з-під*, *з-поміж*, *поміж*, *у ролі*, *у справах*, *на зразок*, *у випадку*, *у порівнянні з*, *посеред*). Властивість використовуватиметься при визначенні пріоритетів при наявності кількох іменників-претендентів на роль ГС.

При визначенні класу *дія* семантичні ознаки виводилися з граматичних ознак слова, заданих його граматичною семантикою. Клас формувався за списком квазіфлексій віддієслівних іменників, що містить кінцеві буквосполучення слів та їх словоформ, за якими в межах граматичних класів іменників можуть визначатися слова на позначення *дії* (*-ання*, *-яння*, *-ення*, *-єння*, *-іння*, *-ація*, *-яція*, *-иття*, *-яття*, *-уття*, *-ництво*, *-ство*, *-ність*, *-илля*, *-овка*, *-обка; -анню*, *-янню* і т. д.)

Але оскільки ці квазіфлексії можуть мати слова з іншою семантикою (*населення*, *інформація*, *поняття*), передбачається перевірка претендентів на роль ГС за електронним семантичним словником прийменникових конструкцій.

Використання блоку правил для конкретних прийменників орієнтоване на неоднозначні ситуації, які репрезентують переважно прийменникове керування при дистантному розташуванні ГС. Сюди ж відносяться правила, які враховують лексико-статистичні характеристики прийменників, а також лексико-семантичні, які передбачають перевірку іменників на входження до семантичних класів (СК) – груп слів, які мають спільну сему в їх лексичних значеннях. Так, для прийменника *в/у* ГС-іменники, які не ввійшли до класів *істота* і *дія* (350 одиниць), було розподілено за 28 семантичними класами. Найбільшими з них є класи *процес* (*переговори*, *побудова*, *рибальство*), *подія* (*арешт*, *вибух*, *виступ*), *рух* (*виїзд*, *експорт*, *перехід*), *захід* (*виставка*, *нарада*, *турнір*), *ставлення/оцінка* (*невіра*, *недооцінка*). Семантичні класи задаються списками квазіоснов. При ідентифікації претендента на роль ГС із квазіосновою зі списку, він отримуватиме мітку головного слова в ЗПЗ.

Синтаксичну розмітку українського речення продемонструємо на прикладі. На вхід синтаксичного аналізатора поступає морфологічно розмічений текст: *У<PF> референдумі<MF> планується<YO> участь<FA> і<SC> палестинців<MI>, які<OS>* ***мешкають<UP> за<PE> межами<FL>*** *автономії<FB> .<e>*. Для зони зв'язків прийменника *за* синтаксична розмітка матиме такий вигляд: *мешкають<UP><10PR> за<PE><11PR> межами<FL><12PR>*, де поруч з морфологічною інформацією приписуватиметься номер залежного слова у реченні і тип синтаксичного зв'язку – *PR* (керування).

У ході встановлення ЗПЗ передбачається розв'язання проблем синтаксичної неоднозначності, а також усунення граматичної омонімії, яка залишилась після АМА.

Література

1. Аполлонская Т. А., Марашлец Е. Ф., Попескул А. Н. Устранение омонимии при АПТ // Инженерная лингвистика и оптимизация преподавания иностранных языков в вузе. – Л.: ЛГПИ, 1983. – С. 92–102.
2. Бугаков О. В. Зони прийменникових зв'язків у синтаксичній структурі українського речення // Мовознавство. – 2005. – № 5. – С. 75–87.
3. Бугаков О. В., Грязнухіна Т. О., Рабулець О. Г. Формування прийменникових текстозорієнтованих реляційних баз даних на корпусі українських текстів // MegaLing'2005 Прикладная лингвистика в поиске новых путей. Материалы международной конференции. 27 июня – 2 июля 2005 Крым, Меганом. – ТНУ им. В. И. Вернадского, Симферополь, 2005. – С. 19–20.
4. Грязнухина Т. А. Анализ предложных связей в научном тексте. – К.: Наукова думка, 1985. – 148 с.
5. Синтаксический анализ научного текста на ЭВМ / Под ред. Т. А. Грязнухиной. – К.: Наукова думка, 1999. – С. 31–32.

Н. Ерреро Бікус*
Київський національний лінгвістичний університет (Київ)

УДК81'322.5–6

АНАЛІЗ І СИНТЕЗ ПЕРФЕКТНИХ ФОРМ АНГЛІЙСЬКОГО ДІЄСЛОВА АКТИВНОГО СТАНУ

The paper focuses on the creation rules of perfect forms for determining paradigmatic classes of English verbs. The material is based on the English-Ukrainian dictionary of the second level provided by the Computational Linguistics laboratory. Had been used the paradigm of English verb, the created program characterizes with division word-change forms into paradigmatic classes of regular (12) and irregular verbs (56). It encapsulates the automatic analysis of perfect forms in the input text and the automatic synthesis of verbs had been taken into account main methods of verb expression.

Відомо, що англійське дієслово характеризується складною і розгалуженою системою словозмінних форм, яка становить значні труднощі для тих, хто вивчає англійську мову. Ця розгалуженість спричинена не так кількістю граматичних категорій, що в англійській мові не більша, ніж в інших індоєвропейських мовах, як специфічним співвідношенням граматичних категорій і засобів їх вираження.

Об'єктом дослідження в роботі є система дієслівних словозмінних форм англійської мови. П**редметом дослідження** – перфектний (доконаний) час англійської мови, правила утворення перфектних форм активного стану. Вибірка дієслів здійснена на основі англо-українського словника ІІ рівня, розробленого в лабораторії комп'ютерної лінгвістики Київського національного лінгвістичного університету. М**ета дослідження** полягає у створенні програми, яка розпізнаватиме в тексті та синтезуватиме перфектні форми дієслів англійської мови.

Перфектні (доконані) часи позначають дії, що закінчилися раніше певного моменту в теперішньому (the Non-Past Perfect) чи минулому (the Past Perfect) і, як правило, пов'язані з цим моментом. Ці часи вважаються вторинними, оскільки виражають дію не саму по собі, а у співвіднесеності з якимось наступним моментом або дією. Доконані часи мають форму активного й пасивного стану, але в нашому дослідженні розглянуто лише дієслова активного стану.

Перфектні часи утворюються з відповідних неозначених часів допоміжного дієслова to have (в Non-Past Perfect (теперішній час) — have, has, це дієслово, яке використовується для утворення аналітичних форм дієслова; в Past Perfect (минулий час) — had) та дієприкметника минулого часу (Past Participle) основного дієслова. В англійській мові Non-Past Perfect ще називають Present Perfect на позначення теперішнього часу, проте майбутнього часу немає, є тільки форми вираження майбутнього часу. Дієслівна парадигма (ДП) — система словозмінних форм дієслова, що виражають граматичні категорії фінітності, способу, стану, часу, числа і особи [1: 13]. ДП в англійській мові досить складна і розгалужена, але точна кількість форм, що її утворюють, залишається