12. Онлайн словник «Академік» [Електронний ресурс]. – Режим доступу: http://dic.academic.ru/

13. Lingvo Online – бесплатный онлайн словарь [Електронний ресурс]. – Режим доступу: http://www.lingvo.ua/ru/Translate/uk-en

Електронний словник синонімів як засіб системного опису синонімічних відношень у лексичній системі української мови

Тетяна Грязнухіна

к. філол. н., старший науковий співробітник, старший науковий співробітник Українського мовно-інформаційного фонду НАН України, Україна, E-mail: ukritaldb@gmail.com

Тетяна Любченко

к. т. н., старший науковий співробітник Українського мовно-інформаційного фонду НАН України, Україна, E-mail: t.lyubch@gmail.com

The article is devoted to the questions of procedural determination of the synonyms with using the theory of the semantic states, to the questions of forming the Electronic dictionary of synonyms (EDS) of the Ukrainian language. EDS is oriented for the using it as an instrument for automatic semantic indexing of the text in the natural language processing systems.

Ключові слова — лексичні синоніми, синсет, електронний словник синонімів, семантичний стан, інтегрована лексикографічна система, комп'ютерні інформаційні технології.

I. Вступ

Синонімія належить до універсальних явищ, властивих усім природним мовам, що вирізняються з-поміж інших семіотичних систем своїм багато-однозначним і одно- багатозначним співвідношенням між позначуваним об'єктом і його позначальним знаком. Синоніми у мові є проявом другого із зазначених типів: різні знаки позначають той самий об'єкт.

Синонімічні відношення завжди були в центрі уваги лінгвістів-дослідників з лексичної семантики, основним положенням (твердженням) якої є декларування високого ступеня системної організації лексичної системи мови, сутність якої проявляється в різноманітних семантичних контекстах – синонімічних, антонімічних, паронімічних, гіпонімічних, без урахування яких здійснення повного опису семантики слова з опертям лише на його референтне значення неможливе. Насамперед це стосується багатозначних слів, доля яких в словнику мови становить біля 40%.

II. Об'єкт і завдання дослідження

Об'єктом даного дослідження є синонімічна підсистема української мови, тобто лексичні синоніми.

Синонімічні відношення порівняно з іншими вказаними вище парадигматичними відношеннями мають більш універсальний характер щодо вияву їх у межах різних лексико-граматичних класів слів. Якщо відношення гіпо- гіперонімії відіграє суттєву роль здебільшого в організації лексичного значення іменника, антонімічні відношення – прикметника, прислівника (градуальні, комплементарні) та дієслова (конверсиви), то синонімічні відношення є базовими в організації значення всіх частин мови.

Загострення уваги дослідників до явища синонімії на сучасному етапі розвитку лінгвістичної науки обумовлено також тією роллю, яку відіграє залучення інформації про синонімічні відношення при розв'язанні завдань автоматичного опрацювання мовної інформації, пов'язаних зі створенням таких нових засобів комунікації, як INTERNET, із завданням підвищення ефективності роботи систем автоматичної обробки текстів (АОТ). В інформаційних системах показник повноти і точності пошуку інформації значно збільшується із включенням до пошукового образу, не тільки ключових слів, але й їхніх синонімів. В системах автоматичного перекладу чітке розмежування перекладних еквівалентів та їхніх синонімічних відповідників призводить до зменшення кількості варіантів перекладу. Супровід перекладних еквівалентів синонімами в автоматичних перекладних словниках підвищує якість перекладу. Синонімічне індексування текстових корпусів забезпечує можливість проведення різних досліджень з питань функціонування синонімів у мовленні на репрезентативному текстовому матеріалі.

В будь-якій інтелектуальній системі АОТ основним джерелом, з якого здійснюється екстракція семантичної інформації, є словник. Є також розуміння того, що в єдиний словник неможливо вмістити всю різнобічну інформацію, необхідну для опису семантики мовної одиниці щодо її лексичного значення, парадигматичних і синтагматичних зв'язків, які є відображенням семантичних контекстів. Як показали результати досліджень з питань комп'ютерної лінгвістики, здійснюваних в Українському мовно-інформаційному фонді під керівництвом академіка НАНУ Широкова А.В., такий багатоаспектний комплексний опис семантики може

ефективно бути реалізованим у межах єдиної інтегрованої лексикографічної системи (ІЛС) (детально теорія описана в [1]). Мовно-інформаційним інструментарієм опису слугують електронні семантичні словники, електронні синтаксичні словники керування як джерело синтаксичної семантики, етимологічні, перекладні словники, інтегровані з базовим в ІЛС електронним словником, як основним носієм інформації про референтне значення слова. Паперовим відповідником останнього є двадцятитомний тлумачний словник української мови.

Основною вимогою до електронних словників є: експліцитне представлення в них інформації у формі, яка була б доступною для використання людиною комп'ютерних засобів пошуку інформації в словнику, а також яка забезпечувала б виконання словником резидентної функції в інших програмах АОТ в ситуаціях, коли користувачем замість людини стає комп'ютер. У даному випадку це означає, що Електронний словник синонімів (ЕСС), з одного боку, повинен виконувати функцію інформаційної системи щодо синонімічної лексики української мови, а з другого, повинен бути спроможним виконувати функцію програмно-лінгвістичного інструмента здійснення автоматичної синонімічної параметризації лексикографічних систем, а також автоматичного семантичного індексування текстів у системах АОТ.

Саме необхідністю забезпечення можливості виконання цих функцій електронним словником синонімів було обумовлено визнання першочерговими завданнями операційне визначення параметрів, за якими встановлюються синонімічні відношення, і розробка зовнішньої і внутрішньої структур ЛБД ЕСС.

III.Операційне визначення параметрів синонімічності

У лінгвістиці загальновизнаним параматром, за яким формується синонімічний ряд (група, гніздо), виступає семантична подібність (близькість, схожість, тотожність) їхніх елементів, яка проявляється в лексичних значеннях цих елементів. При цьому існує безліч інтерпретацій поняття семантичної подібності.

Другим параметром, який визнається більшістю лінгвістів, є властивість синонімів бути взаємозамінними в тексті, яка полягає в тому, шо загальний зміст контексту залишається тим самим при замінюванні в ньому певного слова відповідними йому синонімами. Як і у випадку з першим параметром, для другого також не існує одностайного розуміння, як здійнювати його перевірку.

Слід зауважити, що при визначенні обох параметрів в усіх без винятку дослідженнях в явному чи в неявному вигляді йдеться не про значення слова взагалі, а про його конкретне значення. В тлумачному словнику йому відповідає певне лексичне значення або конкретний відтінок певного лексичного значення.

В ЕСС усе сказане вище знаходить своє експліковане представлення в уточненні поняття лексичних синонімів як лексико-семантичних варіантів (ЛСВ) різних слів, між якими існує семантична схожість. Встановлюване синонімічне відношення повинно задовольняти умові рефлексивності та транзитивності.

Встановлення семантичної подібності ЛСВ різних слів-претедентів на роль синонімів здійснювалося шляхом порівняння відповідних цим ЛСВ значень, зафіксованих у тлумачному словнику (ТЛС), із застосуванням методики, побудованої на теорії семантичних станів, запропонованої А. В. Широковим [2]. Згідно з цією теорією, семантичний стан ЛСВ описується множиною ознак граматичної та лексичної семантики, які задаються за допомогою оператора F, який інтерпретується як оператор сукупності значень певних семантичних категорій: F= {M(S), Z(S), Xi}, де M(S) - ознаки граматичної семантики, що визначають лексико-граматичний клас відповідного слова; Z(S) і Xi - ознаки лексичної семантики, що репрезентують «семантичну тему», яка вказує на референта ЛСВ з оточуючого світу і диференційні ознаки теми. Наприклад, для M(S) = дієслово Z(S) може приймати значення: дія, процес, діяльність, стан (фізичний, ментальний, вольовий). Для Z(S) = «дія» значеннями {Xi} будуть: характер дії, оцінна ознака дії, рух, суб'єкт, об'єкт, адресат, спосіб, засіб виконання, середовище, місце, час, мета, умова, причина, міра, межа, інтенсивність, напрямок. Конкретні значення Xi визначаються в результаті аналізу тлумачення, що відповідає даному ЛСВ. (Детально див. [3, 4]). Семантично схожими визнаються ЛСВ - претенденти на роль синонімів, які мають у формулах описування їхніх семантичних станів однакові значення M і Z, а відповідні їм {Xi} або повністю збігаються, або мають розходження не більше, ніж у два значення.

Вимога до синонімічного відношення бути рефлексивним і транзитивним, висунута в даному дослідженні, знімає існуючу в лінгвістиці неоднозначність у визначенні поняття «взаємозамінність», через те що обумовлює встановлення синонімічних зв'язків для кожного елемента з кожним, наслідком чого є невілювання поняття домінанти (опорного слова) синонімічної групи, а отже, й поняття синонімічного гнізда, коли елементи групи вступають між собою в зв'язки опосередковано через домінанту. З цієї ж вимоги випливає заборона на встановлення синонімічного зв'язку між ЛСВ, референтами яких є родо-видові поняття.

Процедурне визначення параметра взаємозалежності ґрунтується на визнанні залежності, яка встановлена між лексичною сполучуваністю слів і ступенем їхньої синонімічності: чим більший ступінь подібності фіксується в сполучуваності даних слів з іншими, тим більшим є ступінь синонімічності цих слів. Першою спробою формалізувати параметр взаємозамінності при встановленні синонімічних

відношень може вважатися запропонована А.Є. Супруном [5] дистрибутивно-статистична методика визначення семантичної схожості/відмінності слів за показником їхньої лексичної сполучуваності.

Проте, як показали результати аналізу, встановлення лише цієї залежності при процедурній перевірці параметра взаємозамінності синонімів недостатньо, через те що лексична сполучуваність встановлюється на рівні лексем, а синонімічний зв'язок – на рівні їхніх конкретних ЛСВ. І отже, встановлення факту подібності лексичної сполучуваності слів (йдеться про багатозначні слова, до складу яких може входити навіть до 78 ЛСВ) ще не свідчить, що із врахуванням конкретного значення лексем зміст контексту після заміни в ньому даної лексеми на іншу залишиться незмінним. В усякому разі, потрібно передбачити проведення відповідного аналізу, який можна здійснити тільки вручну.

У даному дослідженні перевірка параметра взаємозамінності у ЛСВ– претендентів на роль синонімів здійснювалася із залученням корпусної підтримки як на етапі визначення лексичної сполучуваності претендентів, так й на етапі зіставлення змісту контекстів, за якими визначалася ця сполучуваність. Для цього за допомогою програми семантичної розмітки Українського національного лінгвістичного корпусу, створеного в УМІФ НАНУ, здійснювалося автоматичне маркування синонімів у текстах. За допомогою спеціальної програми за запитом лексикографа для синонімів, які є об'єктом перевірки, автоматично формуються конкорданси заданої довжини (у нашому випадку довжиною в одне речення), з яких виводяться списки лексичної сполучуваності. Суттєва кількісна різниця цих списків, або наявність у них лише одного чи двох спільних компонентів слугували підставою для негативного висновку щодо можливості надання відповідним ЛСВ статусу синонімів через невиконання ними вимоги взаємозамінності. Так, ЛСВ «виконувати» зі значенням «здійснювати щонебудь, реалізувати завдання, наказ, задум і т. ін., проводити в життя» та ЛСВ «нести» зі значенням «виконувати які-небудь обов'язки, доручення і т. ін.», визначені претендентами в синоніми за першим параметром семантичної схожості, після описаної вище процедурної перевірки за другим параметром щодо їх взаємозамінності не були визнані синонімами. Підставою для прийняття такого рішення було те, що за корпусом у відповідних цим ЛСВ списках лексичної сполучуваності було виявлено кількісне розходження в 17 слів (19 – у «виконувати» і 2 – у «нести»), а спільним в списках виявилося одне слово «служба», причому в сполученні з «виконувати» воно зустрілося в одному контексті, а в сполученні з «нести» - в 53 із загального числа 59.[2]

IV.Формування електронного словника синонімів

Базовою інформацією для формування електронного словника синонімів слугувала синонімічна ЛБД, побудована автоматичним шляхом на основі ЛБД комп'ютерної версії двотомного Словника синонімів української мови [6], який містить у собі 9020 синонімічних рядів, поділених на синонімічні групи за допомогою крапки з комою. Вихідну ЛБД було модифіковано в такий спосіб: синонімічні ряди вихідного словника, до складу яких входить кілька груп, в синонімічній ЛБД записано як окремі множини синонімів з перенесенням в них домінанти відповідного ряду; синонімічні ряди дієслів доконаного виду, які у Словнику синонімів задаються через відсилання до відповідних рядів недоконаного виду (відвалити див. відвалювати) записуються в експліцитному вигляді шляхом дублюванням рядів недоконаного виду. В результаті 9020 синонімічних рядів вихідного словника в синонімічній ЛБД ЕСС репрезентовано 15 645 множинами синонімів з вихідного словника, які розглядалися як претенденти для утворення з них синонімічних угрупувань – синсетів. (Назву «синсет» використано на позначення множини слів, взятих у їхніх конкретних значеннях, і які відповідають прийнятим в даному дослідженні параметрам синонімічної схожості й взаємозамінності, через те що назви «синонімічний ряд», «синонімічна група», «синонімічне гніздо» в лінгвістичній літературі термінологізовані з орієнтацією на інше визначення поняття синоніма, відмінне від прийнятого в даному дослідженні).

Статус синсету така множина отримує за умови позитивного результату перевірки її елементів за параметрами семантичної схожості і взаємозамінності, встановлюваними на рівні конкретних значень, зафіксованих в ТЛС, який є найбільш надійним і репрезентативним джерелом інформації про лексичні значення українських слів. При цьому, орієнтуючись на можливість використання ЕСС як резидентної програми, до синсету ставиться вимога експліцитного визначення загального значення, за яким він утворений, і визначення значень його елементів, екстрагованих з ТЛС, а також вимога неможливості входження слова в конкретному значенні в кілька синсетів.

Для забезпечення проведення перевірок в автоматизованому режимі (важко навіть уявити, як це зробити вручну) і для забезпечення автоматизованого формування статей ЕСС за результатами цих перевірок синонімічну ЛБД було інтегровано з ЛБД електронного тлумачного словника на рівні тлумачних статей. В результаті такої інтеграції стало можливим автоматичне встановлення зв'язку для кожного слова синсету із відповідною цьому слову

[2] Уважаємо, що така вузька лексична сполучуваність цього ЛСВ «нести» – практично з одним словом, може слугувати підставою для вилучення його зі складу значень в тлумачній статті лексеми «нести», а словосполучення «нести службу», «нести вахту» записати в тлумачному словнику в зону усталених словосполучень.

тлумачною статтею з ТЛС, представленою для багатозначних слів у вигляді пронумерованих значень (відтінків в середині значення), що інтерпретуються як окремі ЛСВ слова. Зв'язок з тлумачною статтею може здійснюватися як через реєстрове слово ЕСС, так і через будь-який елемент синсету, який аналізується.

Кінцевою метою перевірки елементів синсету є вибір з тлумачної статті для кожного з них конкретного значення (конкретного відтінка певного значення), яке відповідає визначеним параметрам синонімічності, запис вибраного значення і його номера в тлумачній статті у відповідні поля структурованої в ЛБД ЕСС синонімічної статті.

З метою здійснення автоматизованого редагування було розроблено спеціальний формат візуального представлення синонімічних статей (див. рис. 1), при якому в експлікованому вигляді задаються елементи синсету і подається інформація про:

- кількість синсетів, у які входить дане слово (кожний з цих синсетів можна вивести на екран);
- лексико-граматичний клас елементів синсету;
- загальне значення, за яким слова об'єднані в синонімічну групу у вихідному словнику синонімів.

При цьому поле синоніма є активним і дозволяє здійснювати послідовно для кожного елемента синсету зв'язок з ТЛС. Після проведення дослідником аналізу вибране значення автоматично заноситься в поле «значення синоніму» в ЕСС, а також переносяться всі мітки, що супроводжують значення в тлумачному словнику (стилістичні, семантико-синтаксичні) в спеціально відведені для них поля ЛБД

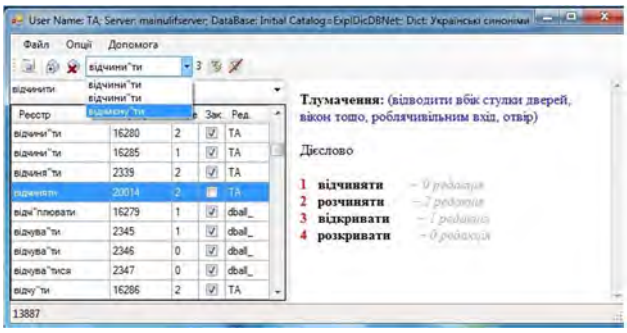

Рис. 1 Візуальне представлення синонімічної статті, яка редагується

Крім надання лексикографу можливості перегляду тлумачної статті з ТЛС, відповідної кожному слову редагованого синсету (див. рис. 2), сервісні програми автоматичного редагування забезпечують:

- автоматичне перенесення вибраного значення і його номера з тлумачної статті в окремі поля синсету в ЛБД ЕСС (номер виконує функцію ідентифікатора ЛСВ, що стає елементом синсету в ЕСС);
- запис значення першого синоніма як загального значення синсету;
- видалення елементів синсету;
- введення нового елемента;
- зміна ієрархічного розташування елементів синсету (при визначенні ієрархії перевага надається ЛСВ однозначного слова та ЛСВ без стилістичних міток);
- видалення синсету;
- формування нового синсету.

Type	Interpretation
ВІДЧИНЯТИ	
Тлум.1	Відводити, відхиляти вбік стулки дверей, вікон і т. ін., роблячи вільним вхід, доступ до чогось або вихід назовні
Відт.1	Робити вільним вхід, доступ до чого-небудь або вихід назовні
Відт.2	Піднімати вгору віко або кришку скрині, коробки і т. ін., роблячи вільним доступ усередину
Тлум.2	<рідко.> Те саме, що [В]відмика́ти[/В] 1
Відт.1	Те саме, що [В]відкрива́ти[/В] 4

Рис. 2 Результат інтегрування синоніма *відчиняти* з його тлумаченням в ТЛС

На рис. 3 зображено кінцевий результат редагування синсету, представленого на рис.1. (відчиняти).

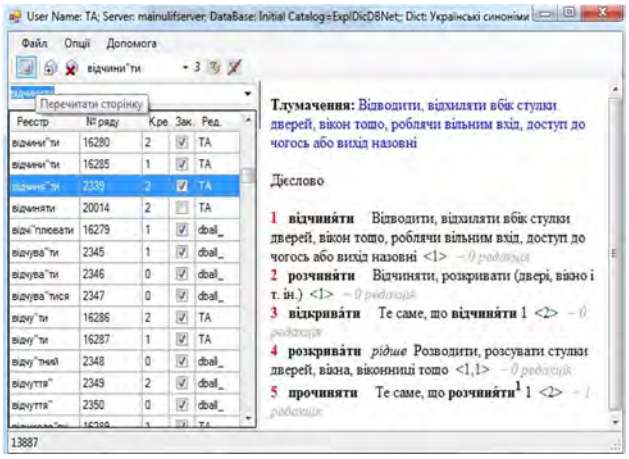

Рис.3. Сформований синсет в ЕСС

Збільшення кількості елементів редагованого синсету (порівн. рис.1 і рис. 3) є наслідком аналізу семи синсетів, утворених з груп синонімів, записаних через крапку з комою в синонімічному ряді «відчиняти» вихідного словника синонімів, на якому формувалася ЛБД ЕСС. В результаті цього аналізу в редагований синсет як раз і було додано ЛСВ «прочиняти<2>», який було вилучено із синсету {відчиняти, прочиняти, прокривати}. В ЕСС крім названого синсету перенесено синсет {відмикати<1>, **відчиняти**<2>, розмикати<2>, відпирати<1>} із загальним значенням «відкривати ключем замок або що-небудь замкнене», а решта п'ять були зруйновані.

Кінцевим результатом проведення в такий спосіб автоматизованого редагування синонімічної бази ЕСС було отримання статусу «синсет» для 13 621 із 15 545 редагованих.

Висновок

Електронний словник синонімів, створений у межах ІЛС, інтегровано в ній з електронним ТЛС за допомогою спеціальних програмних технологій. Внутрішня структура словникових статей забезпечує експліковане представлення інформації про синонімічні відношення в лексичній системі української мови. Все це дає підставу вважати побудований словник ефективним інструментом синонімічної параметризації ТЛС, одним з результатів якої є уточнення лінгвістичного опису значень полісемічних слів додатковою семантичною ознакою – синонімічним контекстом. Крім того, ЕСС є одним з основних засобів семантичної розмітки текстових корпусів.

Автори висловлюють щиру подяку Олені Устимець за її внесок в укладання електронного словника синонімів.

Література

1. Широков В.А. Феноменологія лексикографічниз систем : монографія / В.А. Широков. – К.: Наукова думка, 2004. – 327 с

2. Широков В. А. Семантичні стани мовних одиниць та їх застосування в когнітивній лексикографії / В.А. Широков // Мовознавство. – 2005. – №3 – 4.– С. 47 – 62.

3. Грязнухіна Т. Операційне визначення критерію семантичної подібності синонімів / Т.Грязнухіна, О.Устимець, В.Широков // Прикладна лінгвістика та лінгвістичні технології: MegaLing-2008: Зб. наук.пр. / НАН України, укр. мовно-інформ.фонд та ін..; редкол. : Ю.Д. Апресян та ін.. – К.: Довіра, 2009. – С. 42 – 57.

4. Устимець О. В. Формування електронного словника синонімів української мови / Устимець О. В // Учёные записки Таврического национального университета им. В. И. Вернадского. Серия „Филология". Том 20 (59).– Симферополь, 2007. – № 4.– С. 57 – 61.

5. Общее языкознание / под ред. А. Е. Супрун. – Минск, 1983. – 456 с.

6. Словник синонімів української мови: у 2 т.– К., 1999–2000.

Електронна лексикографічна система «Морфограф»: теоретичні засади та методика конструювання (проект)

Оксана Зубань

к. філол. н., доцент, доцент кафедри української мови та прикладної лінгвістики, Київський національний університет імені Т. Шевченка, Україна, E-mail: oxana.mell.zuban@gmail.com

The article describes theoretical issues, principles of construction and problems set in the project of creating the "Morphograph" electronic lexicographical system representing Ukrainian language morphemics. This system is created on the basis of linguistic morphemic data base containing 200,000 words which applies computer modelling technique of structural-functional connections of morphs in a word. The morphemic data base provides the following benefits: automatic classification of Ukrainian language lexicon according to different parameters of morphemic structure of a word (quantity of morphemes in a word, models of morphemic structures, specified affix or root morph); automatic compilation of root and affix lists on the basis of a dictionary of initial word forms. The "Morphograph" lexicographical system is based on theoretical principles of modern morphemology and sets a goal to systematize and describe basic units of the Ukrainian language morphemic system – morphs, morphemes and morphemic structures of words – in two electronic dictionaries: Dictionary of morphemic structures of words and Dictionary of morphemes. Each dictionary will be compiled in interactive mode according to the user's specified search options and it will also provide the function of copying an automatically compiled list of units.

Ключові слова — комп'ютерна лексикографія, морфемна база даних, морф, морфема, морфна / морфемна структура слова, морфемна парадигма.

Вступ

Українська морфемна лексикографія, започаткована відомими традиційними словниками «Морфемний аналіз: Словник-довідник» І. Т. Яценка [31], «Морфемний словник» Л. М. Полюги [25], із розвитком теорії морфеміки та словотвору поповнилася лексикографічними працями, які ставлять нові лексикографічні завдання і є взірцем словникарства нової доби – комп'ютерної української лексикографії: «Словник афіксальних морфем української мови» [26], «Кореневий гніздовий словник української мови» [17] та «Шкільний словотвірний словник сучасної української мови» [29]. Ці словники нового типу укладалися за допомогою комп'ютера на матеріалі електронної словникозорієнтованої бази даних морфемно-словотвірного фонду Інституту мовознавства ім.О. О. Потебні НАН України, вони існують у традиційному паперовому форматі, а також у форматі комп'ютерних копій паперових словників.

У сучасному комп'ютеризованому світі традиційний «паперовий» словник або його комп'ютерна копія не задовольняють потреб користувача-лінгвіста, який потребує швидкого й ефективного способу добору, оброблення та систематизації лінгвістичної інформації. Тому актуальним завданням у розвитку сучасної української комп'ютерної лексикографії є комп'ютерне моделювання електронних (комп'ютерних)