УДК 81'373.42'42-13:82-32Франко

О. І. Ілик,
Львівський національний університет імені І. Франка, м. Львів

ЗАСАДИ МАРКУВАННЯ ДИСКУРСИВНИХ СЛІВ У КОРПУСІ ТЕКСТІВ МАЛОЇ ПРОЗИ І. ФРАНКА

У статті розглядається методика виявлення та маркування дискурсивних слів (ДС) у корпусі текстів (КТ) малої прози І. Франка. З'ясовано статус ДС. Описаний досвід маркування ДС у корпусній лінгвістиці. З'ясовано, що маркування ДС є елементом прагматичної анотації. Виявлені такі важливі елементи анотування ДС, як вказівка на частиномовну приналежність, розрізнення омонімії, дискурсивного та недискурсивного вживання, маркування прямої та авторської мови, визначення функціонально-семантичних груп, тобто створений повноцінний тегсет для анотування одиниць, які формують структуру дискурсу.

Ключові слова: *дискурсивні слова (ДС), корпус текстів (КТ), маркування, дискурсивна анотація, прагматична анотація, тег, омонімія, пряма мова, авторська мова.*

В статье рассматривается методика выявления и маркировки дискурсивных слов (ДС) в корпусе текстов (КТ) малой прозы И. Франко. Выяснено статус ДС. Описанный опыт маркировки ДС в корпусной лингвистике. Выяснено, что маркировка ДС является элементом прагматической аннотации. Обнаружены такие важные элементы аннотирования ДС, как указание на частеречную принадлежность, разграничение омонимии, дискурсивного и недискурсивного употребления, маркировки прямой и (авторской) речи, определение функционально-семантических групп, то есть создан полноценный тегсет для аннотирования единиц, которые формируют структуру дискурса.

Ключевые слова: *дискурсивные слова (ДС), корпус текстов (КТ), маркирования, дискурсивная аннотация, прагматическая аннотация, тег, омонимия, прямая речь, косвенная (авторская) речь.*

This article deals with methods of detecting and marking discursive words (DW) in text corpus of Ivan Franko's small prose. The author makes an attempt to clarify the status of DW. The experience of marking-up of DW in corpus linguistics is covered. It was found that marking-up of DW is a part of pragmatic annotation. The author of the article identifies the following essential elements of DW-annotation: the indication of parts of speech, distinguishing of homonyms, discursive and nondiscursive use, marking-up of direct and author's speech, the definition of functional-semantic groups, ie a proper tag-set to annotate units that form the structure of discourse.

Key words: *discourse words (DW), text corpus, mark-up, discursive annotation, pragmatic annotation, tag, homonymy, direct speech, author's speech.*

XX ст. – це період розквіту науки про мову, час, коли з'являються нові синтезовані дисципліни. Чільне місце серед них займають корпусна лінгвістика та лінгвістична прагматика. Стаття ґрунтується на теоретичних засадах цих двох напрямів. Тому **актуальність теми** зумовлена першою спробою їхнього поєднання у межах одного дослідження: на основі КТ малої прози І. Франка запропоновано здійснити маркування ДС, які, в свою чергу, належать до сфери лінгвістичної прагматики. ДС є одними з найчастотніших у творчості І. Франка, мають велике функціональне навантаження і важливі для створення словника мови письменника.

Наукова новизна дослідження полягає в тому, що це перша спроба перенести дані, які є на матеріалі текстів, на рівень образності, на рівень прагматики, на рівень, так би мовити, живого життя слова.

Незважаючи на широкий науковий інтерес до проблеми дискурсивних слів, немає загальноприйнятої дефініції. По-перше, відсутня усталена термінологічна база для позначення цих елементів. У лінгвістичній традиції найчастотнішою є терміносполука «дискурсивні слова» [1; 2; 3; 9]. Проте їх часто називають **модальними словами** [6]; **прагматичними маркерами** [15]; **дискурсивними маркерами** [14; 19; 22; 23**]** і т. д. По-друге, нема чіткого переліку одиниць, які мають бути включені до розряду ДС. Тому існує багато підходів (традиційний, функціонально-семантичний, комунікативний та ін.) до вивчення ДС. Такі елементи належать до різних частин мови (частки, вставні та модальні слова, прислівники, вигуки, сполучники, прийменники, деякі класи займенників), але об'єднані вони спільною функцією: формування структури дискурсу та забезпечення зв'язності тексту.

Важливо зазначити, що ДС – це одиниці не структури мови, а мови у функціонуванні (ДС стають дискурсивними лише у процесі функціонування): «у комунікації вони функціонують завдяки не стільки принципам взаємодії частин мови, скільки принципам уживання, «живого життя» слова в процесах інтеракції особистостей. Інакше кажучи, ДС – категорія комунікативна, а значить – антропна, людиноцентрична» [2, с. 248]. Тому важливими є контексти вживання ДС, ситуативні значення.

Оскільки віднесення того чи іншого слова до розряду дискурсивних переважно залежить від контексту вживання і семантичного та прагматичного навантаження у ньому, тому маркування ДС у КТ значно полегшить пошук ДС, допоможе встановити їх арсенал.

З метою забезпечення точного та всеохоплюючого аналізу ДС у межах художнього дискурсу важливо провести процедуру маркування ДС.

Вона передбачає виконання таких завдань: 1) визначення принципів маркування ДС у практиці деяких КТ; 2) створення повнотекстового електронного варіанту творів малої прози І. Франка; 3) прикріплення тегів до кожного ДС; 4) зняття омонімії; маркування частини мови; 5) маркування прямої / непрямої мови; 6) маркування функціонально-семантичних груп ДС.

1) Маркування ДС у практиці деяких КТ

Анотація є основною характеристикою корпусу і саме присутність такої інформації відрізняє його від простих колекцій чи бібліотек, адже основна мета електронних бібліотек – зміст текстів, а не їх мовні осо-

бливості. Під лінгвістичною анотацією у корпусній лінгвістиці традиційно розуміють: а) довільну лінгвістичну інформацію про лінгвально-релевантні одиниці текстових даних; б) практику введення формалізованої лінгвістичної інформації в електронний текст; в) наявність такої інформації у тексті» [8]. Поряд із поняттям «**анотування**» використовують терміни «**розмітка**», «**індексація**» (індексування), «**маркування**».

У КТ розрізняють **зовнішнє анотування** (дані про автора: стать, вік, освіта; дані про твір: час написання, виходу друком, редакцію) та **внутрішнє** – внутрішньомовна інформація (анафорична, просодична, семантична, фонетична, дискурсивна, прагматична, частиномовна чи морфолого-синтаксична анотація).

Із розвитком таких наук, як лінгвістична прагматика та дискурс-аналіз, популярними, але не до кінця дослідженими, стали дискурсивне та прагматичне маркування КТ. Оскільки ДС формують структуру дискурсу і є носіями найтонших прагматичних смислів, тому важливо визначити, до якого виду анотації належатиме їх маркування. Розглянемо ці два види маркування детальніше.

Чіткого визначення терміносполуки «дискурсивне анотування» нема. **Дискурсивну розмітку**, як правило, використовують для анотування корпусів усного (розмовного) мовлення. Зокрема, у Корпусі японської розмовної мови (The Corpus of Spontaneous Japanese / Nihongo hanashikotoba koopasu) [21], а також у корпусі під назвою «Рассказы о сновидениях и другие корпуса звучащей речи» [10], в основі якого – дискурсивне анотування даних живого російського мовлення. Синонімом до дискурсивної розмітки у цьому випадку є дискурсивна транскрипція, в основі якої лежать такі поняття, як елементарні дискурсивні одиниці, дискурсивні маркери, просодичні характеристики мовного потоку, прискорена вимова і т. д.

Прагматичній анотації, звичайно, присвячено менше досліджень ніж іншим видам розмітки. Це пов'язано з тим, що прагматичну розмітку роблять напівавтоматичним способом і вручну. Особливістю прагматики є те, що значення залежать від контексту, тому основне завдання прагматичної анотації – в повній мірі враховувати контекст.

Прагматичну анотацію часто включають до так званої трирівневої структури дослідження: синтаксична, семантична і прагматична анотації [7; 18], проте вона є і об'єктом окремих наукових розвідок [13; 19]. Бувають випадки, коли прагматичне маркування сплутують з анафоричною анотацію та кореферерeнційними зв'язками. Хоча визначення анафоричних зв'язків також відносять до сфери прагматики. [18]. Крім того, визначення та позначення ДС також є складовою прагматичної анотації [19]. Проте, є й окремі праці присвяченні маркуванню ДС у КТ [14; 19; 22; 23].

У рамках спільного проекту російських і французьких науковців, присвяченому аналізу часток, вийшла книга «Путеводитель по дискурсивным словам русского языка» [1], в якій на основі машинних корпусів текстів створили корпус прикладів вживання ДС. Це дає можливість читачу перевірити функціонування слова на більш широкому мовному матеріалі. Саме корпус прикладів завершує словникову статтю ДС, яке інтерпретується.

Отже, дослідження ДС є надзвичайно популярними, вони дедалі частіше привертають увагу сучасних дослідників [1; 2; 9].

2) Створення повнотекстового електронного варіанту творів малої прози І. Франка

У Львівському університеті існує проект створення корпусу текстів Івана Франка [5]. Передбачається, що корпус буде охоплювати всі твори І. Франка (а це близько 7 млн слововживань). Створення корпусу текстів ДС малої прози Івана Франка є частиною цього проекту. Уся мала проза І. Франка займає вісімнадцять збірок оповідань та новел. Основою для нашого аналізу є 17 збірок україномовних оповідань (≈ 437533 тис слововживань).

За допомогою КТ ДС малої прози І. Франка стане можливим отримати якісну і кількісну характеристику ДС. А в перспективі укласти частотний словник (ЧС) та конкорданс ДС малої прози І. Франка.

3) Теги для позначення ДС

У лінгвістичній традиції використовують різні теги (спеціальні коди, які через формальний запис експлікують граматичні значення слів, до яких вони приписані [8]) на позначення ДС.

Так, наприклад, G. Leech і M. Weisser [17] розробили схему анотації мовленнєвого акту, в якій виділяють окремий тег на позначення дискурсивних маркерів <dm>(**d**iscourse **m**arker).

D. Samy і A. González-Ledesma розробили теги для прагматичного маркування – PRAGMATEXT [19]. Відповідно до схеми, прагматична інформація маркується тегом <PI> (**p**ragmatic **i**nformation). Його можна використовувати для маркування різних прагматичних елементів на синтаксичному рівні і для анотації дискурсивних маркерів.

Для маркування ДС у КТ малої прози І. Франка пріоритетним буде тег <dw> (**d**iscourse **w**ords), який, як і терміносполука «дискурсивні слова», підкреслюватиме специфіку відповідних елементів як слів, які «співвідносять зміст висловлюваного з комунікативною ситуацією дискурсу, відсовуючи на другий план їхні формальні характеристики і наголошуючи на порівнянні із мовним використанням, а не мовною системою чи структурою» [2, с 9].

4) Частиномовне маркування ДС: розрізнення омонімії

ДС – це, перш за все, одиниці дискурсу, для яких важливо виявляти комунікативні смисли, а не частиномовну приналежність. Проте частиномовна анотація є необхідним етапом маркування ДС. Вона допоможе профільтрувати слова-омоніми, які не мають прагматичного навантаження (випадки, коли ДС є просто дієсловами чи іменниками), відкинути недискурсивне вживання ДС.

Для морфологічного маркування ДС використано символи, в яких великі латинські літери позначають відповідну частину мови: V (**V**erb), N (**N**oun), D (A**d**verb), AR (**A**uxiliary p**r**eposition), AP (**A**uxiliary **p**article).

У маркуванні ДС виникають труднощі через те, що майже всі ДС формують омонімні зв'язки зі словами інших лексико-граматичних класів, тобто є міжчастиномовними омонімами. Тому розрізнення омонімії серед ДС є необхідним, позаяк «у багатьох словах такого типу існують, поряд з дискурсивними, й інші, недискурсивні вживання» [9, с. 9].

Наприклад, порівняймо вживання слова *може:*

(1) Ані хреста, ані надгробного каменя на своїй могилі не веліла класти. «Життя дало мені все, що могло дати, нехай же й смерть бере все, що ***може****<V|МОГТИ> взяти», – говорила вона. («Батьківщина»)*[*20]

*(2) Се його син! А проте він поводився з ним, як з ворогом, як з чимось чужим, ненависним. Не міг інакше! <dw>****Може#****<A##|МОЖЕ></dw>, з часом привикне, полюбить його, але тепер він не міг, не міг дивитися на нього, не міг знести дотику його дитинячої руки, його рожевих уст! Не міг та й годі! («Odi profanum vulgus»)*

У першому контексті слово *може* – дієслово, яке виражає процесуальність (може / могти), а другий контекст – це дискурсивне вживання. Слово набуває іншого значення – «невпевненості мовця у сказаному».

У корпусі текстів малої прози І. Франка омонімію розрізнено шляхом контекстного аналізу, додавши до слова відповідні позначки «#»: може – могти (дієслово) V; може# – може (вставне слово) А##.

5) Маркування прямої / непрямої мови

У статті важливо ввести поняття «режим інтерпретації». Так, за О. В. Падучевою, є два режими інтерпретації: 1) мовленнєвий, комунікативний (діалогічний) режим інтерпретації; 2) наративний режим інтерпретації. Відповідно і ДС в діалозі та наративі різні. В діалогах переважають ДС: *ну, авжеж, бігме*, а в наративі ДС: *правда, справді, звичайно, отже*.

У лінгвістичній науці маркування прямої та непрямої мови здійснено в Французькому розмовному корпусі (French Oral Narrative Corpus) [16], корпусі англійських діалогів 1560–1760 [12], у корпусі творів М. Достоєвського[11] та в корпусі текстів І. Франка [4].

Маркування прямої та авторської мови належить до структурного типу розмітки. У міжнародному стандарті кодування текстової інформації (TEI) [20], зокрема в розділі про цитування, використовують тег <said> на позначення прямої мови. Він присутній і в маркуванні прямої мови в малій прозі І. Франка. Тег також включає вказівку на мовця (допоможе визначити, які ДС уживають ті чи інші персонажі). Наприклад:

– Е, ні, я на таке не пишуся. Кажіть, подобалася вам чи ні?

*<said speaker=«Опанас»>– <dw>****Авжеж****<AP|АВЖЕЖ></dw> сподобалася.</said> («Батьківщина»)*

Для маркування ДС, яке знаходиться в авторському мовленні, пропонуємо тег <narrator>:

Побіч нього жили якісь панство і мали служницю, сільську дівчину, зовсім непоказну з лиця.<narrator><*dw*>***Отже***<A##|*ОТЖЕ*></*dw*>*, доля хотіла, що тої самої ночі панство були на балу, а дівчина ночувала сама в кухні.*</narrator> («Odi profanum vulgus»)

6) Маркування функціонально-семантичних груп ДС

Нині є багато класифікацій ДС. Найпопулярніші – В. Виноградова [6], А. Баранова, В. Плунгяна, К. Рахіліної [1], Ф. Бацевича [2], В. Белової [3], Б. Фрейзера [15], Д. Самі, А. Гонсалес-Ледесми [19].

Для дослідження ДС у корпусі текстів малої прози І. Франка використано функціонально-семантичну класифікацію, яка є найпоширенішою. Оскільки ДС – це одиниці не структури мови, а мови у функціонуванні, тому ця класифікація є пріоритетною. Адже для ДС важлива їх функція в дискурсі, а лексико-семантичний аналіз не допоміг би розкрити смисли ДС.

Тому при маркуванні ДС доцільно вказувати групу, до якої належить ДС. Так у Д. Саммі і А. Гонзалес-Ледесми [19] при анотації ДС використовують позначку [DR] (**D**iscursive **R**elations), яка вказує на групи дискурсивних маркерів.

Для маркування ДС у корпусі текстів малої прози І. Франка використаємо позначку [DF] (**D**iscursive **F**unction). Відповідно до груп ДС, розрізняють такі функції ДС: впевненість (*звичайно, авжеж, розуміється*) / невпевненість (*певно, видно, мабуть, здається*); узагальнення (*загалом, отже, взагалі*) / перерахування (*далі, по-перше, по-друге*); поєднання (*і, або*) / уточнення (*власне*); емоційність (*на щастя, на жаль, прикро*); суб'єктивація (*чую*) / об'єктивація (*кажуть*); допустовість (*все-таки, одначе*) та ін.

Отже, повністю промарковане ДС за вищеперерахованими параметрами матиме такий вигляд:

*<narrator><dw DF=«емоційність»>****На щастя****<A##|НА_ЩАСТЯ></dw>, мама вийшли були до городу, а то би, певно, були насварили на мене, нащо беру такий великий кусень хліба.</narrator>(«Микитичів дуб»)*

*<said speaker=«Тоньо»>– А <dw DF=«невпевненість»>****чень****<AP|ЧЕНЬ></dw> же він і сам троха піде?</said> – сказав Тоньо. («Ґава і Вовкун»)*

Висновок

Дискурсивна лексика перебуває в центрі уваги багатьох наукових досліджень, проте в статті вперше застосовано корпусний підхід до їх виявлення та аналізу.

Запропонована схема анотування ДС у корпусі текстів малої прози І. Франка дасть змогу встановити арсенал ДС, визначити частотність ДС у творах письменника та відповідно – чисельність функціонально-семантичних груп ДС; ДС, які характерні для мовленнєвого та наративного режиму інтерпретації, а порівняння отриманих частот з ЧС сучасної української художньої прози допоможе визначити особливості ідіолекту письменника.

Розрізнення омонімії у корпусі ДС малої прози І. Франка доповнить Словник омонімів української мови та СУМ, а найголовніше – контексти вживання цих слів можуть бути використані з дидактичними цілями у навчальному процесі.

---

[*] Тут і далі подаємо приклади з корпусу текстів малої прози Івана Франка.

Опрацювання всіх контекстів вживання ДС здається неможливим без використання новітніх комп'ютерних технологій. Саме контексти визначають смислове навантаження ДС, тому створення корпусу ДС малої прози І. Франка є єдиноможливим шляхом цілісного комплексного опрацювання ДС у творчості письменника. Описане анотування значно полегшить мовознавчі та літературознавчі дослідження. Розглянуті елементи маркування можна вважати необхідними як для дослідження ДС у творчості І. Франка, так і інших письменників.

Література:

1. Баранов А. Н. Путеводитель по дискурсивным словам русского языка / А. Н. Баранов, В. А. Плунгян, Е. В. Рахилина. – М. : Азбуковник, 1993. – 243 с.

2. Бацевич Ф. С. Частки української мови як дискурсивні слова: монографія / Ф. С. Бацевич. – Львів : ПАІС, 2014. – 288 с.

3. Белова В. М. Функционально-семантические особенности дискурсивных слов в жанре мемуаров / В. М. Белова // Вестник Череповецкого государственного университета. – 2011. – № 1. – С. 50–53.

4. Бук С. Н. Пряма й авторська мова великої прози Івана Франка: лінгвостатистичне дослідження у контексті корпусної лінгвістики / С. Н. Бук // Вісник Львівського університету. Серія філологічна. – 2011. – Вип. 52. – С. 199–209.

5. Бук С. Н. Корпус текстів Івана Франка: спроба визначення основних параметрів / Бук С. Н. // Прикладна лінгвістика та лінгвістичні технології: MegaLing-2006: Зб. наук. пр. / НАН України. Укр. мовн.-інформ. фонд, Таврійськ. нац. Ун-т ім. В. І. Вернадського; за ред. В. А. Широкова.– К. : Довіра, 2007. – С. 72–82.

6. Виноградов В. В. Русский язык (грамматическое учение о слове) / В. В. Виноградов. – М., 1947. – 783 с.

7. Гладкова Г. П. Модель анотації текстового корпусу як засіб дослідження художньої картини світу / Г. П. Гладкова // Studia Linguistica. – Випуск 4. – 2010. – С. 524–528.

8. Демська-Кульчицька О. М. Базові поняття корпусної лінгвістики / О. М. Демська-Кульчицька // Українська мова. – № 1. – 2003. – С. 40–46.

9. Дискурсивные слова русского языка: Опыт конкретно-семантического описания / Под ред. К. Киселевой и Д. Пайара. – М. : Метатекст, 1998. – 447 с.

10. Рассказы о сновидениях и другие корпуса звучащей речи. – [Цит. 22 марта 2014]. – Доступно з <http://www.spokencorpora.ru/>.

11. Шайкевич А. Я. Статистический словарь языка Достоевского / А. Я. Шайкевич, В. М. Андрющенко, Н. А. Ребецкая. – М., 2003.

12. A Corpus of English Dialogues 1560–1760 – Uppsala universitet. – [Cited 18 March 2014]. – Available from : <http://www.engelska.uu.se/Research/English_Language/Research_Areas/Electronic_Resource_Projects/A_Corpus_of_English_Dialogues/>.

13. Archer, Dawn, Culpeper, Jonathan and Matthew Davies. Pragmatic annotation // Corpus linguistics: an international handbook / edited by Anke Ludeling and Merja Kyto. Mouton de Gruyter. – 2008. – P. 613–642.

14. Darinka Verdonik, Matej Rojc, Marko Stabej Annotating discourse markers in spontaneous speech corpora on an example for the Slovenian language // Language Resources & Evaluation; May 2007. – Vol. 41 Issue 2. – P. 147–180.

15. Fraser B. Pragmatic markers // Pragmatics 6(2). – 1996. – P. 167–190.

16. French Oral Narrative Corpus. – [Cited 18 March 2014]. – Available from : <http://frenchoralnarrative.qub.ac.uk/>.

17. Leech G., Weisser M. Generic Speech Act Annotation for Task-Oriented Dialogue // Archer/Rayson/Wilson/McEnery (eds.) 2003. Proceedings of the Corpus Linguistics 2003 Conference. Lancaster University : UCREL Technical Papers, vol. 16. – 2003. – P. 441–446.

18. Navarro B., Civit M., Martí A., Marcos R., Fernández B. Syntactic, Semantic and Pragmatic Annotation in Cast3LB. Corpus Linguistics 2003 Workshop on Shallow Procesing of Large Corpora. UCREL Technical Report, Lancaster (UK), 2003. – P. 59–68.

19. Samy D., González-Ledesma A. Pragmatic Annotation of Discourse Markers in a Multilingual Parallel Corpus (Arabic-Spanish-English) // LREC 2008, Marrakech, may 2008. – P. 3299–3305.

20. TEI : Text Encoding Initiative. P6 : Elements Available in All TEI Documents. – 2004. – [Cited 18 March 2014]. – Available from : <http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/CO.html>.

21. The Corpus of Spontaneous Japanese. – [Cited 15 March, 2014]. – Available from: <http://www.ninjal.ac.jp/corpus_center/csj/misc/preliminary/index_e.html>.

22. Zufferey S., Popescu-Belis A. Towards Automatic Identification of Discourse Markers in Dialogs: The Case of 'Like'. In: Michael Strube and Candy Sidner (Ed.). SIGdial 2004 (5th SIGdial Workshop on Discourse and Dialogue). Cambridge (Mass., USA). [s.l.]: ACL – Association for Computational Linguistics, 2004. – P. 6–71.

23. Zufferey S., Popescu-Belis A. Annotation of discourse markers (like, well) in the ICSI-MR corpus. – 2004. – [Cited 18 March 2014]. – Available from : <http://www.issco.unige.ch/en/research/projects/im2/mdm/data/discourse-markers/index.html>.