ЛІНГВІСТИЧНІ ТЕХНОЛОГІЇ МОДЕЛЮВАННЯ МОВНОЇ СИСТЕМНОСТІ

В.А.Широков, д.т.н.

НОВА КЛАСИФІКАЦІЯ УКРАЇНСЬКИХ ДІЄСЛІВ

Запропоновано нову класифікацію українських дієслів

The new classification of the Ukrainian verbs is proposed

У мові виявлено не так вже й багато об'єктивних законів, які репрезентуються точними, математичного ґатунку моделями. Аналогічне твердження стосується і класифікацій, створених на основі вивчення одиниць мовних систем. Взагалі побудова кожної нової моделі, яка була б строгою у математичному значенні цього поняття, є подією досить рідкісною. Наскільки ми можемо судити, досвід українського мовознавства не дає підстав для великого оптимізму щодо побудови формальних лінгвістичних моделей, тому оголошення про створення такої теорії для певного класу мовних явищ або одиниць мабуть звучить дещо інтригуючо.

Завдання, яке ми ставимо у цій статті, полягає у тому, щоби показати, що існує точна і цілком формальна – підкреслимо це – класифікація українських дієслів, джерелом якої є їхні граматичні властивості. Оскільки для читача може здатися малоймовірною побудова строгої та ще й нової класифікації у такій, здавалося б, добре вивченій галузі як граматика дієслова, почнемо, так б ми мовити, з формулювання "перших принципів", яких ми дотримуватимемося у цій статті і які у даному випадку є такими.

За традицією природознавства коректно побудована теорія повинна задовольняти трьом основним вимогам:

1) Вона мусить базуватися на мінімально можливому числі аксіом (або постулатів). Припущення ad hoc є недопустимими;

2) Теорія повинна пояснювати усі відомі на момент її створення явища, які належать до розглядуваного класу;

3) Теорія мусить мати певну передбачувальну силу і правильно прогнозувати та пояснювати нові явища або факти.

Наскільки нам відомо, у мовознавчій традиції немає практики висунення до теорії таких жорстких вимог, і може здатися, що вони є занадто сильними для того, щоб їм могла задовольнити будь-яка змістовна і життєздатна теоретична схема. Спробуємо довести, що це не так.

Автором ще понад десять років тому було побудовано основи інформаційної теорії лексикографічних систем, головні результати якої викладено в книзі[1] (надалі – ІТЛС) і апарат якої послідовно застосовувався до розбудови формалізму цілої низки лексикографічних праць й насамперед – структурної теорії 11-томного Словника української мови (надалі – СУМ)[2].

В ІТЛС побудовано формальні репрезентанти структури лівої та правої частин словникових статей СУМа. Зроблено це було у спосіб, який має внутрішні потенції до узагальнень – це надало можливість для твердження про існування буквально єдиної формули, яка описує всю структуру СУМа. Розвинутий підхід залишав можливість і для детального опису та розгляду структур окремих елементів 11-томника. Зокрема, для структур лівих частин дієслівних словникових статей було одержано репрезентанти у вигляді певних математичних формул, які мали просту і однозначну інтерпретацію як графи певного типу.

При побудові класифікації українського дієслова ми виходимо з трьох аксіом, які навіяні аналізом структури лівих частин дієслівних словникових статей СУМа, але, як відзначалося у дисертації автора[3], правомірним є й більш абстрактний погляд на лексикографічні структури СУМа – вони мають право на самостійне існування, тобто вже не як факти словника, а взагалі як факти мови, у зв'язку з чим наводилися приклади і пропозиції щодо їхнього можливого застосування у галузях, досить далеких від лексикографії.

Перша аксіома є цілком очевидною:

(1) Кожне українське дієслово реалізується в мові лексемою з визначеною і зафіксованою семантикою, якій притаманні одне або два значення категорії виду.

Наприклад, одне значення виду мають дієслова:

"АХАТИ" (недок.) – недоконаний вид;

"БУХИКНУТИ" (док.) – доконаний вид;

"АБСОЛЮТИЗУВАТИ" (недок. і док.) – має форми як недоконаного так і доконаного виду[4] і т.д.

Двома значеннями виду характеризуються дієслівні лексеми:

“ВДИХАТИ” (недок.) і “ВДИХНУТИ” (док.);

“ПРОСІВАТИ” (недок.) і “ПРОСІЯТИ” (док.) і т.д.

Лексикографічне представлення приналежності лексеми до певного значення атрибута “ВИД” (як воно подано, наприклад, в 11-томнику) будемо далі називати *видовим комплексом* або просто *комплексом*[5]. Наприклад, у словниковій статті:

ЗАСТИЛА́ТИ, а́ю, а́єш *і* **ЗАСТЕЛЯ́ТИ**, я́ю, я́єш *і рідко* **ЗАСТЕ́ЛЮВАТИ**, юю, юєш, *недок.*, **ЗАСЛА́ТИ**, стелю́, сте́леш *і* **ЗАСТЕЛИ́ТИ**, стелю́, сте́лиш, *док.*, *перех.*

наявні два видові комплекси (позначатимемо їх символи C_1 та C_2):

C_1 = **ЗАСТИЛА́ТИ**, а́ю, а́єш *і* **ЗАСТЕЛЯ́ТИ**, я́ю, я́єш *і рідко* **ЗАСТЕ́ЛЮВАТИ**, юю, юєш,

C_2 = **ЗАСЛА́ТИ**, стелю́, сте́леш *і* **ЗАСТЕЛИ́ТИ**, стелю́, сте́лиш,

причому елементи з C_1 об’єднуються у видовий комплекс ремаркою *недок.*, а елементи з C_2 належать до іншого комплексу – з ремаркою *док*.

Друга аксіома не така очевидна:

(2) Кожна дієслівна лексема з визначеною і зафіксованою семантикою та конкретним значенням виду може реалізуватися дієсловами, що належать не більше як до трьох різних словозмінних (парадигматичних) класів.

Для лексикографічного представлення цього факту будемо далі вживати термін *парадигматичний блок*, або просто *блок*[6].

Наприклад, в комплексі C_1 вищенаведеної словникової статті з реєстровим словом “**ПРОВО́ДИТИ**” наявні три блоки – B_{11}, B_{12} та B_{13}:

B_{11} = **ЗАСТИЛА́ТИ** – об’єднуються у блок ремаркою: а́ю, а́єш;

B_{12} =**ЗАСТЕЛЯ́ТИ** – об’єднуються у блок ремаркою: я́ю, я́єш;

B_{13} = *рідко* **ЗАСТЕ́ЛЮВАТИ** – об’єднуються у блок ремаркою: юю, юєш.

Нарешті третя аксіома:

(3) Кожна дієслівна лексема з конкретним значенням виду, визначеною приналежністю до конкретного словозмінного класу та визначеною й зафіксованою семантикою може реалізуватися дієсловами, які мають не більше чотирьох фонетичних варіантів (як правило, це префіксальна та/або коренева варіація).

Лексикографічне представлення цього факту, так само як і в ІТЛС, позначатиметься терміном *компонент*.

Наприклад, в словниковій статті:

ЗСИХА́ТИ *і рідко* **ІЗСИХА́ТИ,** а́ю, а́єш, *недок.*, **ЗСО́ХНУТИ** *і рідко* **ІЗСО́ХНУТИ, ЗСО́ХТИ** *і рідко* **ІЗСО́ХТИ,** хну, хнеш; *мин. ч.* зсох *і* зсо́хнув, ла, ло; *док.*

наявні два комплекси, кожний з яких має по одному блоку, причому в першому комплексі C_1 та його єдиному блоці B_{11} наявні два компоненти:

K_{111} = **ЗСИХА́ТИ** ; K_{112} = **ІЗСИХА́ТИ**;

у другому комплексі C_2 та його єдиному блоці B_{21} наявні вже чотири компоненти:

K_{211} = **ЗСО́ХНУТИ**; K_{212} = **ІЗСО́ХНУТИ**; K_{213} = **ЗСО́ХТИ**; K_{214} = **ІЗСО́ХТИ.**

Можна дискутувати щодо обґрунтованості, розумності, доцільності і т. ін. наведених аксіом, але не можна не погодитися із тим, що вони є формально визначені на множині дієслів, для яких встановлено вищенаведені лінгвістичні параметри. Отже, у цьому випадку маємо справу з формально визначеною системою.

Розвинемо математичний формалізм для опису викладених лінгвістичних фактів. Спочатку розгляд вестиметься для дієслів, що характеризуються одним значенням виду. Уведемо деякі позначення.

Символом I^1 позначимо множину чисел 1, 2, 3, 4:

$$i \in I^1,\ i = 1, 2, 3, 4, \qquad (1)$$

елементами якої нумеруватимемо дієслова-компоненти, визначені 3-ою аксіомою – тобто фонетичні варіанти дієслів, що характеризуються конкретним значенням словозмінних параметрів; таких фонетичних варіантів, як ми постулювали, може бути від 1 до 4. Тоді, якщо застосувати разом усі три аксіоми, то будь-який з елементів множини (1) визначатиме певний видовий комплекс. А саме: задання числа i (i = 1, 2, 3, 4) сигналізує, що маємо дієслова, які характеризуються одним значенням атрибута “ВИД” з притаманним їм одним набором значень парадигматичних показників, а кількість фонетичних варіантів точно дорівнює i. Очевидно, що декартовий добуток $I^1 \times I^1 \equiv I^2$ параметризує структури двоблочних комплексів, а $I^1 \times I^1 \times I^1 \equiv I^3$ – структури триблочних комплексів. Взагалі одержуємо картину, представлену в Додатку 1.

У такий спосіб структура будь-якого видового комплексу представляється однією з трьох сигнатур:

$$(i),\quad (i\ j),\quad (i\ j\ k), \qquad (2)$$

у кожній з яких числа i, j, k незалежно пробігають множину 1, 2, 3, 4. Сигнатурі (i) відповідає комплекс з одним парадигматичним блоком, який має i компонентів; сигнатура $(i\ j)$ представляє комплекс, який має два парадигматичні блоки, у першому з яких наявні i, а другому – j компонентів; сигнатура $(i\ j\ k)$ представляє комплекс, який має три парадигматичні блоки, у першому з яких наявні i, у другому – j , а у

третьому – *k* компонентів. Очевидно, що одноблочних сигнатур може бути максимум 4, двоблочних – 16, триблочних – 64. У такий спосіб видовий комплекс може реалізовуватися максимум 84-ма різними структурами (4 + 16 + 64 = 84).

Словникова стаття, яка містить обидва видові комплекси, зображається парою (α . β), де α і β незалежно пробігають множини сигнатур {(*i*)}, {(*i j*)}, {(*i j k*)}, поданих у формулі (2). Очевидно, що сигнатур типу (α . β) може бути максимум 7056 (84×84).

Отже, загальне число структур, які задовольняють аксіомам (1) – (3), теоретично може дорівнювати 7140 (84 + 7056). Позначимо їх через: (α) – однокомплексні сигнатури та (α . β) – двокомплексні сигнатури. Графічна інтерпретація сигнатур (α) і (α . β) аналогічна наведеній в ІТЛС.

У процесі створення нової версії СУМа в Українському мовно-інформаційному фонді була сформована комп'ютерна лексикографічна база даних 11-томника, з використанням якої було проведено обчислювальний експеримент із визначення саме тих сигнатур (α) й (α . β), що насправді реалізовані у лексикографічній системі цього словника. Такий експеримент було здійснено співробітниками Фонду Н.М.Сухариною та К.М.Якименком на усьому масиві СУМа шляхом автоматичного аналізу його лексикографічних структур. У результаті на масиві 41402 дієслівних словникових статей із потенційно можливих 7140 класів лівих частин було виявлено всього 52 класи[7]. Їхній список наведено у Додатку 2.

Ми не ставимо собі тут за мету проведення ґрунтовного лінгвістичного аналізу збудованої системи. Зробимо лише декілька зауважень.

По-перше, викладена система визначає певну класифікацію на множині українських дієслів. Справді, позначимо одержані класи дієслів через λ_1, λ_2, ..., λ_{52} , а множини дієслів, що належать до відповідного класу – через $q(\lambda_1)$, $q(\lambda_2)$,..., $q(\lambda_{52})$. Тоді очевидно, що:

$$q(\lambda_i) \cap q(\lambda_j) = \varnothing \quad \text{при } i \neq j \,, \qquad (3)$$

тобто будь-яке дієслово може належати до одного і тільки до одного з визначених вище класів[8]. Це означає, що одержана нами класифікація є коректною.

Крім того, вона є доволі точною – показово, що з неї не знайшлося жодного винятку на масиві понад 41 тисяча дієслівних лексем. Отже, можна із достатньою упевненістю стверджувати, що одержана

класифікація інкорпорує усі факти з розглядуваного класу і у такий спосіб репрезентує певний об'єктивний закон української мови.

Декілька слів про "порожні" структури класифікації, тобто ті, для яких не знайшлося підходящих дієслів в СУМі. Нас цей факт не бентежить. Згадаймо, наприклад, історію із побудовою Д.І.Мєндєлєєвим класифікації хімічних елементів – число клітин у ній теоретично взагалі не обмежувалося, хоча у ті часи були відомі не більше як 63 елементи, та ще й чимало з них були не дуже добре ідентифіковані. У створеній Мєндєлєєвим таблиці був цілий ряд прогалин, але система – періодичний закон – дозволяла прогнозувати властивості "пропущених" елементів[9].

Так само і в нашому випадку не виключене існування певних дієслів, які належатимуть до "пропущених" класів. Наявність вільних класів сигналізує про досі не використані словотвірні потенції українського дієслова. Можливо, більш глибокі дослідження та розвиток мовної системи дозволять виявити такі класи, тим більше, що система сама "прогнозує" їх морфологічні властивості. Так, серед однокомплексних сигнатур наявні: (11) – наприклад: **БУЛЬКОТА́ТИ,** очу́, о́чеш *і* **БУЛЬКОТІ́ТИ,** очу́, оти́ш, *недок.*, (12) – наприклад: **РИБА́ЛИТИ,** лю, лиш, **РИБА́ЛЧИТИ** *і рідко* **РИБА́ЧИТИ,** чу, чиш, *недок,* але вже сигнатура (121) – відсутня. З нашої класифікації випливає, що сигнатура (121) визначає видовий комплекс з трьома парадигматичними блоками, у першому з яких наявний 1, у другому – 2 і у третьому – 1 компонент. Отже для елементів цього класу, за аналогією з варіантом сигнатури (12), сигнатурі (121) може відповідати така модель лівої частини словникової статті:

Х+А́ЛИТИ, лю, лиш, **Х+А́ЛЧИТИ** *і рідко* **Х+А́ЧИТИ,** чу, чиш,
Х+А́ЧУВАТИ, ую, уєш, *недок* ,

де символом **Х** позначено корінь гіпотетичної лексеми.

З іншого боку, інтуїтивно зрозуміло, що реалізація "надважких" класів, які репрезентуються сигнатурами, наприклад (344.444), (444.344), (444.444), дуже малоймовірна – їх просто "не витримає" морфологічна система української мови. Отже, постають питання про встановлення меж цієї системи та виявлення нових закономірностей будови українського дієслова, відповіді на які можуть бути одержані на шляху проведення комплексних лексико-граматичних і лексико-семантичних досліджень з використанням одержаних структур формальної класифікації українського дієслова та лексикографічних баз даних Українського мовно-інформаційного фонду.

На закінчення автор висловлює подяку Н.М.Сухарині та К.М.Якименку за співробітництво, а також В.М.Русанівському, який погодився прочитати рукопис цієї статті і зробив цінні зауваження та поради.

[1]Широков В.А.Інформаційна теорія лексикографічних систем. –К., 1998. − С. 331; [2]Словник української мови в 11 томах. – К., 1970 – 1980; [3]Широков В.А. Інформаційна теорія та системотехнічні засади комп'ютерної лексикографії. Дис...Д-ра техн. наук. – К., 1999; [4]Зауважимо, що категорія "ВИД", розглядувана як інформаційний атрибут, в СУМі набуває одного з шести значень, а саме: "недок.", "док.", "недок. і док.", "док. і недок.", "недок. і рідко док." та "док. і рідко недок.". Отже, будь-який дієслівний комплекс характеризується одним з наведених значень; [5]В ІТЛС цій структурі відповідає термін *блок*; [6]В ІТЛС йому відповідає термін *підблок*. [7]Фактично спочатку було ідентифіковано 53 типи і серед них один, який не вкладався у розбудовану нами систему, а саме: тип із сигнатурою 1111 (тобто – чотириблочною, що суперечило 2-й аксіомі). У цьому класі містилася лише одна стаття із лівою частиною "**ЗАСТРУМЕНІ́ТИ,** заструмени́ть *і* заструмені́є *і* **ЗАСТРУМИ́ТИ,** и́ть *і* **ЗАСТРУМІ́ТИ,** і́є *і* **ЗАСТРУМУВА́ТИ,** у́є, *док.*". Проведений Н.М.Сухариною аналіз показав, що семантика членів реєстрового ряду не є тотожною, а це суперечить аксіомам (1–3). Унаслідок цього аналізу члени реєстрового ряду були розведені за окремими словниковими статтями. Більш докладному викладові цього факту буде присвячено окрему працю. Показово, що збудована нами система вже на першому кроці її застосування виявила свою діагностичну силу; [8]Класифікаційною для дієслів є й повна множина 7140 класів з сигнатурами (α) та $(\alpha \,.\, \beta)$, оскільки і для них справедлива формула (3); [9]Навіть у наш час відкрито тільки 105 елементів, причому останні з них, надважкі, є надзвичайно нестійкими (мають дуже малий час життя), в природі не спостерігалися, а були одержані лише штучним шляхом.