**Костіков Микола Павлович**
**Костиков Николай Павлович**
**Mykola Kostikov**

# ПІДХІД «РОЗУМНИХ ПАРАДИГМ» ДЛЯ СТВОРЕННЯ БАЗИ ЗНАНЬ МОРФОЛОГІЇ ПОЛЬСЬКОЇ МОВИ

**Анотація.** У статті аналізуються сучасні підходи до формалізації морфології польської мови та можливості їх використання при створенні електронних засобів навчання іноземних мов. Розглядаються особливості моделювання мови з допомогою так званих «розумних парадигм».

**Ключові слова:** бази знань, граматика, експертні системи, електронні засоби навчання, моделювання мови.

# ПОДХОД «УМНЫХ ПАРАДИГМ» ДЛЯ СОЗДАНИЯ БАЗЫ ЗНАНИЙ МОРФОЛОГИИ ПОЛЬСКОГО ЯЗЫКА

**Аннотация.** В статье анализируются современные подходы к формализации морфологии польского языка и возможности их использования при создании электронных средств обучения иностранным языкам. Рассматриваются особенности моделирования языка с помощью так называемых «умных парадигм».

**Ключевые слова:** базы знаний, грамматика, моделирование языка, экспертные системы, электронные средства обучения.

# ‘SMART PARADIGM’ APPROACH FOR CREATING A KNOWLEDGE BASE OF POLISH LANGUAGE MORPHOLOGY

**Annotation.** The article is concerned with modern approaches to Polish morphology formalization and possibilities of using them for creating computer-aided language learning software. Language modeling with «smart paradigms» is analyzed.

**Keywords:** CALL, e-learning, expert systems, grammar, knowledge bases, language model.

**Постановка проблеми в загальному вигляді.** Нині у світі розробляється велика кількість електронних засобів навчання (ЕЗН) іноземних мов, у тому числі і для польської. Проте і вибір, і якість програмних засобів для вивчення слов'янських мов і досі недостатні. Зокрема бракує ЕЗН польської мови, у яких був би ґрунтовно реалізований зворотний зв'язок і враховувались особливості рідної мови користувача [1]. У той же час викладання слов'янських мов в Україні має свою специфіку, оскільки вони високофлективні та споріднені до української, а відповідно більшого значення при вивченні набуває засвоєння граматики. Виправити ситуацію з ЕЗН польської мови можна шляхом створення експертно-навчальної системи. Такі засоби дозволяють реалізувати ефективний зворотний зв'язок у процесі навчання.

**Аналіз останніх досліджень і публікацій.** Принципи побудови інтелектуальних навчальних і експертно-навчальних систем (ЕНС) досліджували Ю. Машбиць, О. Меняйленко, В. Петрушин, Д. Смолін та ін. Питанням формалізації граматики польської мови займалися зокрема М. Волінський, Р. Волош, В. Ґрущинський, І. Новак-Коморовська, З. Салоні, А. Сляський, Я. Токарський.

**Виділення невирішених раніше частин загальної проблеми.** Важливими складовими експертних систем є база знань (БЗ), що розв'язує задачі в певній предметній області, та підсистема пояснень, яка дозволяє користувачеві прослідкувати хід розв'язання й пересвідчитись у обґрунтованості кожного кроку [2, с. 24]. Тому для створення ЕНС граматики польської мови спершу слід змоделювати предметну область таким чином, аби забезпечити прозорість і зрозумілість дій, які система виконує при розв'язанні відповідних задач, зокрема при виборі та утворенні граматичних форм (ГФ) слова. Формалізація граматики в тому чи іншому вигляді реалізована в численних програмних засобах. Морфологічний синтез є важливим завданням у таких застосуваннях обробки природної мови, як електронна лексикографія і машинний переклад, однак моделювання граматики для цих цілей не враховує специфіки, притаманної процесу вивчення іноземної мови. Для створення ЕНС необхідно розробити

таку БЗ із правилами граматики, зокрема морфології, яка б не просто моделювала синтез ГФ слів, а й забезпечувала наочність і можливість пояснення цього процесу.

**Мета статті** — проаналізувати сучасні підходи до моделювання граматики польської мови і визначити складові БЗ морфології, необхідні для синтезу ГФ при створенні відповідної ЕНС.

**Виклад основного матеріалу.** Якомога більш повне моделювання словозміни флективних мов є головною задачею електронних граматичних словників, що розробляються для багатьох мов, у тому числі й для польської. У 2007 р. у Варшаві було випущено перше, а в 2012 р. — друге видання електронного граматичного словника польської мови SGJP (Słownik gramatyczny języka polskiego). Основою цього та інших подібних засобів традиційно є реляційна модель, у якій слова класифікуються і групуються за частинами мови та типами відмінювання, і в окремих таблицях зберігаються набори закінчень ГФ для кожного з типів. Як зазначає співрозробник SGJP М. Волінський, ця уніфікована і відносно компактна модель дозволяє врахувати всі тонкощі польської словозміни [3, с. 96]. Однак поділ слів на велику кількість парадигматичних класів у подібних моделях ускладнює роботу користувача з ними, а також не дає можливості побачити в явному вигляді закономірності утворення конкретних ГФ, що потрібно при дослідженні та вивченні мови. За словами А. Ранта, у традиційній парадигматичній моделі немає точного визначення парадигми та її застосування [4, с. 16].

Граматика польської мови моделюється також при розробці засобів морфологічного аналізу. Зокрема при створенні програми Morfologik у відповідність списку ГФ польської мови було поставлено граматичні значення, які вони виражають. Отже, у подібних засобах теж певним чином моделюється словозміна, вже без прив’язки до парадигматичних класів. Однак і тут утворення ГФ лише описується, а не пояснюється. Крім того, в існуючому вигляді ці засоби дозволяють лише аналізувати, а не синтезувати ГФ.

Словник для перевірки орфографії ispell містить файл із описом шаблонів словозміни та словотвору. З його допомогою можна згенерувати всі можливі ГФ для слів, що містяться у словнику. Кожне слово має один чи декілька «прапорців» — міток, яким у допоміжному файлі відповідають правила утворення цілої парадигми чи її частини. Через регулярні вирази описано формальні умови для автоматичної розмітки

прапорцями початкових форм слів. Шаблони супроводжуються коментарями та прикладами слів, які підпадають під описані зразки. Проте в кожному шаблоні наводиться лише перелік усіх нетотожних ГФ без розмітки за граматичними значеннями, які вони виражають.

При роботі над багатомовним перекладачем європейських мов MOLTO, що триває з 2010 р., для моделювання словозміни співрозробники застосували так звані «розумні парадигми», які спершу було використано при дослідженні фінської морфології А. Ранта [5, с. 130]. Суть цього підходу полягає в тому, що існує лише одна загальна функція утворення всіх ГФ для кожної частини мови, але ця функція має певну кількість аргументів. Система аналізує одну чи декілька заданих ГФ і після цього сама обирає необхідний повний шаблон утворення парадигми слова з можливих для відповідної частини мови [6, с. 646]. Таким чином, завдяки опису загальних залежностей немає необхідності наперед знати парадигматичний клас конкретного слова, оскільки більшість закономірностей у словозміні можна визначити, аналізуючи безпосередньо саме слово. Більш того, застосувавши правила, що містяться в цій моделі, теоретично можна згенерувати парадигму будь-якого слова (в тому числі неологізму), якщо тільки буде зазначено його частину мови. На відміну від засобів із іншими моделями формалізації граматики, процес роботи «розумних парадигм» при словозміні більш схожий на дії експерта, що має відповідні знання та може робити деякі припущення.

При виборі шаблону словозміни система послідовно зіставляє введене слово з умовами, яким мають відповідати ті чи інші його літери. Порядок запису і проходження переліку цих умов суттєвий. Спочатку слово перевіряється на часткові випадки, далі — на більш загальні. Для прикладу, при дієвідмінюванні в англійській мові правильні дієслова, що закінчуються на *e*, мають спільні правила утворення парадигми, крім випадків із закінченнями *ie* та *ee*. Тому спершу перевіряються дві останні літери, і лише якщо слово не підпадає під ці часткові випадки, відбувається перевірка одного останнього символу на значення *e* [7, с. 24]. Подібним чином реалізується утворення парадигм для інших частин мов.

Описаний підхід використовується при створенні програмних засобів із обробки природної мови, і його застосування автоматизовує роботу лексикографів [6, с. 645], тобто людей, що є експертами в тій чи іншій мові. Під час поповнення електронного словника для них зручніше ввести дві-три ключові ГФ слова, ніж шукати в довіднику

номер одного з десятків чи сотень парадигматичних класів. Однак деякий відсоток невизначеності все ж лишається, і остаточний висновок щодо коректності парадигми, відтвореної за цією спрощеною процедурою, все одно робить людина, яка знає мову. Іноді для точної ідентифікації шаблону словозміни необхідно ввести 4–5 ГФ. В експерименті, проведеному для іменників і дієслів англійської, шведської, французької та фінської мов, коректність парадигми, реконструйованої за однією ГФ, варіювалась від 76% до 97% залежно від мови та частини мови [6, с. 650].

Натомість для людини, що тільки-но починає вивчати мову, недостатньо такого механізму утворення ГФ слів, який лише частково спрощує цю задачу і потребує при цьому допомоги користувача. Ефективна ЕНС повинна вміти сама давати точні та однозначні відповіді на поставлені завдання, а також пояснювати їх. Отже, ми пропонуємо видозмінити та доповнити описаний підхід «розумних парадигм» таким чином, аби він відповідав задачі вивчення граматичних правил студентами і міг бути використаний при розробці ЕНС граматики.

Спершу розглянемо ті елементи підходу, які були б корисними для моделювання словозміни при створенні такої ЕНС. По-перше, це аналіз закінчення і основи слова при виборі шаблону словозміни (на відміну від його вибору за міткою парадигматичного класу). Правила, що базуються на таких зовнішніх ознаках слова, до певної міри відображають реальний процес прийняття відповідного рішення людиною-експертом. Багато з цих правил у явному вигляді описані в підручниках із граматики. Другим важливим корисним моментом у «розумних парадигмах» є виділення правил морфологічних і фонетичних перетворень у окремі функції, що дозволяє уникнути дублювання опису тих самих явищ для різних частин мови [7, с. 28]. Це стосується, наприклад, чергувань літер, що було виділено розробниками у спеціальні функції, записані окремо від шаблонів словозміни.

Ці два елементи в підсумку дають на виході більш компактний, чіткий і зрозумілий опис граматичних правил, а тому їх доцільно використати і при створенні ЕНС граматики. Тепер визначимо, що необхідно змінити або додати до вищевказаних елементів підходу «розумних парадигм» при формалізації граматики для створення ЕНС на основі цієї моделі.

По-перше, було б доцільно розділити шаблони утворення цілої парадигми на правила утворення окремих ГФ. У нинішньому вигляді «розумні парадигми» фактично є

надбудовою над вищезгаданою традиційною моделлю з поділом слів на парадигматичні класи. У цій моделі для кожного класу зазначаються правила утворення всіх можливих ГФ. Тим часом окремі класи часто різняться між собою лише одним чи декількома правилами, а решта інформації фактично дублюється. Одним зі шляхів уникнення цього дублювання може бути використання об'єктно-орієнтованої моделі та наслідування спільних ознак між класами. Проте це рішення є спрощенням лише з технічної точки зору і саме по собі не розв'язує проблеми наочного представлення правил. У таких моделях не аналізуються загальні закономірності утворення окремих ГФ і взаємозв'язки між ними, а словозміна описується лише з чіткою прив'язкою до конкретних парадигматичних класів. Отже, розділення парадигми на окремі ГФ дасть можливість і позбутися надлишковості, і більш явно представити правила граматики.

По-друге, корисним буде змінити аргументи функції утворення ГФ таким чином, аби ще більш наблизити модель до реальної ситуації прийняття рішення людиною. У словниках зазвичай зазначається початкова форма та значення класифікаційних граматичних категорій слів. Наприклад, практично в усіх словниках польської мови вказується рід іменника. Інші ж ГФ слова, які б могли бути використані при виборі шаблону словозміни «розумними парадигмами», рідше наводяться у словниках і з меншою ймовірністю відомі студентам. Такі форми радше самі по собі можуть бути об'єктом вивчення. Отже, при утворенні ГФ доцільним буде використовувати початкові форми та інформацію про класифікаційні граматичні категорії слів.

Наприклад, для польського іменника на вибір правил утворення ГФ того чи іншого слова впливають такі характеристики, як його рід (в орудному відмінку: *koc* (ч. р.) → *kocem, noc* (ж. р.) → *nocą*) і категорія істоти / неістоти (у знахідному відмінку: *kot* (істота) → *kota, płot* (неістота) → *płot*). Однак і цих ознак буває недостатньо. Як справедливо зауважує А. Сляський, у польській мові багато прикладів часткової омонімії, коли з початкової форми неможливо згенерувати правильну словозміну іменників, навіть зазначивши граматичний рід [8, с. 31–32].

У ході аналізу навчальної літератури з граматики польської мови та подальшого наповнення БЗ морфологічними правилами нами було визначено зокрема такі додаткові ознаки, що можуть бути важливими для вибору правил в окремих випадках: походження слова — питомо польське чи іноземне (у давальному відмінку: *ziemia* (польське) → *ziemi, akademia* (грецьке) → *akademii*); так звана конкретність / абстрактність (у родовому

відмінку: bal («колода») → bala, bal («бал») → balu). Ці та інші ознаки слів, як і традиційні мітки частин мови, так само можна зберігати у БЗ, а їх використання також уподібнює процес синтезу ГФ системою до дій людини.

Таким чином, при виборі шаблону словозміни для поданого слова система буде керуватися: 1) аналізом початкової форми; 2) класифікаційними граматичними категоріями; 3) додатковими ознаками. Всі ці правила, що визначають вибір закінчень і суфіксів, називатимемо власне морфологічними правилами. Вони застосовуватимуться на першому етапі синтезу ГФ. Натомість на другому етапі, коли вже визначено необхідні морфологічні складові ГФ, задіюватимуться фонологічні правила, котрі, як було сказано вище, доцільно записувати окремо від морфологічних. Вони відповідають за коректну інтеграцію визначених компонентів і стосуються відображення на письмі перетворень, що відбуваються, наприклад, при зміні закінчення слова (пом'якшення *k*: *jabłko* + *em* → *jabłkiem*, позначення м'яких звуків перед голосними: *koń* + *a* → *konia* тощо).

Як морфологічні, так і фонологічні правила, що використовуються при синтезі ГФ, повинні мати коментарі та за потреби демонструватися користувачеві. Ці коментарі входитимуть до складу підсистеми пояснень, яка є необхідним компонентом ЕНС. Загалом розробка та використання такої моделі дозволяє створювати, розв'язувати та пояснювати навчальні вправи з морфології.

Подібний підхід, як і універсальні «розумні парадигми», в подальшому можна застосувати також і до інших флективних мов. Його технічна реалізація можлива з використанням функційних або об'єктно-орієнтованих мов програмування, а інтерфейс навчальної оболонки для простої та зручної роботи з ЕНС доцільно розробити як веб-додаток.

**Висновки та перспективи подальших досліджень.** Отже, підхід до формалізації граматики, реалізований у «розумних парадигмах», може бути корисним при створенні ЕЗН іноземних мов. Із деякими видозмінами його можна успішно застосувати зокрема для синтезу ГФ слів при розробці ЕНС, що ґрунтуватиметься на моделі мови та надаватиме системні знання про застосування правил граматики. Предметом подальшого дослідження є розробка інших складових ЕНС.

## Література

1. *Костіков М. П.* Можливості сучасних електронних засобів навчання польської мови для вивчення граматики студентами / Микола Костіков // Наукові записки. Серія : педагогічні науки / Кіровоград. держ. пед. ун-т ім. В. Винниченка. – Кіровоград : РВВ КДПУ ім. В. Винниченка, 2012. – Вип. 108. – Ч. 1. – С. 206–209.
2. *Петрушин В. А.* Экспертно-обучающие системы / Петрушин В. А. ; отв. ред. А. М. Довгялло ; АН УССР. Ин-т кибернетики. – К. : Наукова думка, 1992. – 196 с.
3. *Woliński M.* A Relational Model of Polish Inflection in Grammatical Dictionary of Polish / Marcin Woliński // Human Language Technology: Challenges of the Information Society. – Berlin, Heidelberg : Springer-Verlag, 2009. – P. 96–106.
4. *Ranta A.* How Predictable is Finnish Morphology: An Experiment in Lexicon Construction : CLT Seminar, 25 September 2008 [Електрон. ресурс] / Aarne Ranta. – 2008. – 64 p. – Режим доступу : www.cse.chalmers.se/~aarne/talks/finnish-2008.pdf.
5. *Ranta A.* How Predictable is Finnish Morphology? An Experiment on Lexicon Construction / Aarne Ranta // Resourceful Language Technology: Festschrift in Honor of Anna Sågvall Hein / Edited by Joakim Nivre, Mats Dahllöf and Beata Megyesi. – Uppsala : University of Uppsala, 2008. – P. 130–148.
6. *Détrez G.* Smart Paradigms and the Predictability and Complexity of Inflectional Morphology : Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, April 23–27, 2012 / Grégoire Détrez and Aarne Ranta. – Avignon, 2012. – P. 645–653.
7. *Ranta A.* Creating Linguistic Resources with the Grammatical Framework / Aarne Ranta. – Valetta : ELRA, 2010. – 75 p.
8. *Slaski A.* Opis fragmentu języka polskiego w formalizmie Grammatical Framework : praca magisterska na kierunku informatyka / Adam Slaski. – Warszawa, 2010. – 73 p.