УДК 004.9:811.162.1

Костіков М. П.

Асистент кафедри інформаційних систем Національного університету харчових технологій, Київ, Україна

ФОРМАЛЬНА МОДЕЛЬ СЛОВОЗМІНИ ІМЕННИКІВ ПОЛЬСЬКОЇ МОВИ

При створенні електронних засобів навчання актуальним питанням є формалізація знань для їх наочного подання та спрощення їх обробки. Формальний опис граматики, зокрема словозміни, наявний у сучасних засобах обробки природної мови, лише фіксує існуючі граматичні форми слів, однак не аналізує процесу їх утворення, що необхідно для цілей вивчення мови.

У статті запропоновано підхід до формалізації словозміни польської мови (на прикладі іменників) шляхом виділення найпростіших перетворень у процесі словозміни. Проведено відбір слів згідно з частотним словником польської мови. Сформовано еталонну таблицю граматичних форм для визначеного набору слів. Описано окремі перетворення при словозміні іменників, і з їх допомогою проведено моделювання процесу утворення словоформ із еталонної таблиці. Одержані в результаті моделювання ланцюжки перетворень для кожного слова збережено в базі даних.

Отримані ланцюжки дозволяють крок за кроком описати утворення граматичних форм слів. За рахунок цього забезпечується наочність подання процесу словозміни для студентів, що вивчають мову. Крім того, можливий автоматичний підбір слів, у яких відбувається те чи інше граматичне явище, для формування навчальних вправ, прикладів і тестових завдань.

Ключові слова: граматика, електронні засоби навчання, комп'ютерна лінгвістика, модель мови, морфологія, обробка природної мови.

НОМЕНКЛАТУРА

БД – база даних;

ГФ – граматична форма.

ВСТУП

Сучасна інформатизація освіти вимагає розроблення нових підходів до процесу навчання. Аби повною мірою розкрити потенціал інформаційних технологій, змін має зазнати не лише форма подання, а й внутрішня структура навчального матеріалу. Знання в електронних засобах навчання повинні бути формалізовані таким чином, щоби зробити їх більш наочними, а побудову окремих елементів навчання – легшою. Це можна реалізувати за рахунок детальної декомпозиції предметної області, виділивши базові елементи знання та відношення між ними.

Для польської мови, вивчення якої набуває в Україні дедалі більшої популярності, досі бракує електронних засобів навчання, які б забезпечували набуття системних знань щодо застосування граматичних правил. Польська, як і інші флективні мови, має багату словозміну, а тому при її вивченні особливої ваги набуває граматика. Тож актуальним є питання, як краще формалізувати знання в цій предметній області. Розроблення навчальної системи, що базується на моделі мови, дозволило би зробити вивчення граматики простішим та наочнішим, а керування процесом навчання – гнучкішим і ефективнішим.

1 ПОСТАНОВКА ЗАДАЧІ

Метою розроблення моделі словозміни польської мови є покроковий опис цього процесу в зрозумілому та наочному вигляді. Також повинна забезпечуватись можливість автоматичного підбору слів, у яких при словозміні відбуваються ті чи інші явища, для генерації навчальних вправ, прикладів і тестових завдань із граматики.

2 ОГЛЯД ЛІТЕРАТУРИ

Проблему формалізації граматики природних мов досліджували А. Залізняк, А. Ранта [1], І. Шевченко, В. Широков [2] та ін. Питанням формалізації граматики польської мови займалися зокрема М. Волінський [3], Р. Волош, В. Ґрущинський, І. Новак-Коморовська, З. Салоні [4], А. Сляський, Я. Токарський.

При моделюванні граматики флективних мов, до яких відноситься і польська, логічним є почати цей процес із формального опису морфології, або словозміни. Інформація про утворення різних ГФ слів на сьогоднішній день доступна для багатьох мов, для яких розроблено електронні граматичні словники та засоби морфологічного аналізу. Зокрема існує ряд моделей, які у формальному вигляді описують словозміну польської мови. Вони дозволяють ефективно розв'язувати задачі морфологічного аналізу та синтезу. Тим не менше, існуючі зараз моделі морфології польської мови, які використовуються у граматичних та орфографічних словниках і морфологічних аналізаторах, не пристосовані до цілей навчання мови. Для опанування граматики іноземної мови необхідно не просто завчити певну кількість словоформ, а й розуміти суть того, що відбувається при їх утворенні. Наявні ж моделі лише описують, а не пояснюють процес словозміни.

В електронних граматичних словниках для різних мов можна побачити опис морфології, формалізований відповідно до задач цих програмних засобів. У подібних лексикографічних системах інформація про словозміну традиційно зберігається у реляційних базах даних. В основу реляційної моделі покладено поділ слів на парадигматичні класи. Слова класифікуються та групуються за типами відмінювання, і в окремих таблицях (за частинами мови) зберігаються набори закінчень ГФ для кожного з типів. Реляційна модель була використана зокрема при створенні граматичного словника української мови [2], а також при створенні граматичного словника польської мови SGJP, автори якого прагнули до повного моделювання польської словозміни [4]. Реляційна модель дійсно дозволяє врахувати й описати всі тонкощі словозміни польської мови, будучи при цьому уніфікованою та відносно компактною [3], проте поділ слів на велику кількість парадигматичних класів утруднює роботу з моделлю, а також не дає можливості побачити законо-

DOI 10.15588/1607-3274-2015-4-3

мірності утворення тих чи інших ГФ. Співрозробники засобу обробки природної мови PoliMorf, які при створенні свого ПЗ використовували SGJP в якості джерела даних, також відзначають як його ваду те, що внутрішня організація даних у цьому словнику є досить складною [5]. Взагалі, за словами А. Ранта, у традиційній парадигматичній моделі немає точного визначення самої парадигми та її застосування [1]. Таким чином, моделювання граматики з виділенням парадигматичних класів за зовнішніми ознаками словозміни, на якому базуються сучасні електронні лексикографічні системи, не враховує специфіки, притаманної процесу вивчення іноземної мови.

Засоби морфологічного аналізу для польської мови (наприклад, [6]) містять списки всіх словоформ для десятків і сотень тисяч слів разом із мітками їхніх граматичних значень (відмінок іменника, час дієслова тощо) і дають змогу працювати з великими обсягами слів та відповідних ГФ. Однак пояснення самого процесу їх утворення тут також відсутнє.

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.

Подібна ситуація спостерігається і при формалізації граматики інших флективних мов. Основними задачами при моделюванні словозміни є морфологічний аналіз і синтез для машинного перекладу, пошуку інформації в мережі Інтернет тощо. Для ефективного розв'язання цих завдань достатньо лише фіксації кінцевих ГФ слів без аналізу процесу їх утворення. Через це розглянуті моделі важко використовувати у процесі навчання у зв'язку з його специфікою.

3 МАТЕРІАЛИ І МЕТОДИ

З метою наочного подання та пояснення процесу словозміни польської мови нами було розпочато розроблення нової моделі. При моделюванні словозміни було вирішено декомпозувати цей процес до якомога простіших елементів, виділивши так звані елементарні перетворення – окремі зміни в написанні слова при утворенні ГФ, такі як додання, вилучення або заміну літери чи послідовної групи літер. Комбінації цих перетворень повинні дозволяти утворити будь-яку ГФ слова від початкової.

Моделювання словозміни іменників польської мови через елементарні перетворення проводилось наступним чином. Спершу було визначено початковий обсяг слів, словозміну яких моделюватиме наша система. З метою подальшої перевірки точності моделі, що розробляється, було створено еталонну таблицю (табл. 1), де для кожного слова мають зберігатися всі його ГФ. БД було наповнено словоформами іменників, витягнутими автоматично (з допомогою розробленого алгоритму) зі словника морфологічного аналізатора польської мови Morfologik [6].

Далі було досліджено процес утворення похідних ГФ іменників від початкових. На основі інформації з підручників і граматичних довідників польської мови було проведено декомпозицію цього процесу з метою виділення елементарних перетворень.

Моделювання процесу словозміни проходило ітераційно в напівавтоматичному режимі. Спершу формально описувались окремі елементарні перетворення. Після цього через спеціально розроблений інтерфейс додання ланцюжків перетворень проводилися спроби змоделювати утворення ГФ слова від початкової. Якщо введена вручну послідовність перетворень приводила до утворення ГФ, записаної в еталонній таблиці, такий успішний ланцюжок додавався до БД. Якщо ж перетворень, описаних досі, було недостатньо, описувались нові перетворення, після чого спроби моделювання повторювались. При наступних ітераціях система спочатку перебирала вже наявні в БД ланцюжки, і лише в разі неуспіху відбувалося введення послідовності перетворень вручну.

Таким чином було отримано ланцюжки перетворень для всіх ГФ визначеного набору слів. Ланцюжки було також записано в БД до окремої таблиці.

4 ЕКСПЕРИМЕНТИ

При визначенні початкового набору іменників для моделювання словозміни було використано частотний словник [7], з якого вибрано 1000 найбільш частотних слів польської мови. Серед них виявилось 356 іменників.

Інформацію про явища, що відбуваються при утворенні ГФ, було взято з підручників польської мови, зокрема [8] і [9]. Елементарні перетворення було описано як окремі методи мовою програмування C# із використанням регулярних виразів.

ГФ слів, інформацію про елементарні перетворення та їхні ланцюжки було збережено з допомогою СУБД SQLite.

5 РЕЗУЛЬТАТИ

У підсумку для визначеного набору іменників було отримано 75 перетворень, які позначено та згруповано наступним чином (табл. 2).

У табл. 3 наведено збережені в БД ланцюжки перетворень для слова «*życie*» при утворенні різних відмінків однини. Перетворення T002 відповідає відкиданню закінчення *-e*, а T006, T012, T024 – доданню закінчень *-a*, *-em*, *-u* відповідно.

Таблиця 1 – Фрагмент еталонної таблиці ГФ іменників

ID	SgN	SgG	SgD	...	SgV	PlN	...	PlV
353	żal	żalu	żalowi	...	żalu	żale	...	żale
354	żołnierz	żołnierza	żołnierzowi	...	żołnierzu	żołnierze	...	żołnierze
355	żona	żony	żonie	...	żono	żony	...	żony
356	życie	życia	życiu	...	życie	życia	...	życia

Таблиця 2 – Елементарні перетворення для 356 найчастотніших іменників польської мови

Тип	Група	Елементарні перетворення
1	відкидання закінчення	–a, –e, –ę,–o, –um
2	додання закінчення	+a, +ą, +ach, +ami, +e, +ego, +em, +emu, +ę, +i, +im, +mi, +o, +om, +oma, +owi, +owie, +ów, +u, +y, +yma
3	додання суфіксу	+ci, +ni, +on
4	випадіння голосної	–(-e-)
5	вставка голосної	+(-e-), +(-y-)
6	чергування звуку або послідовної групи звуків	a→e, ą→ę, c→cz, ch→sz, cześć→ czć, ćc→c, d→dź, dszcz→dżdż, dz→ż, dźń→dń, el→oł, ę→ą, ę→e, g→dz, g→ż, k→c, k→cz, ł→l, o→e, o→ó, ó→o, r→rz, rz→r, s→ś, sł→śl, sn→śń, st→ść, t→ć, z→ź, zn→źń, –(-i-), +(-i-)
7	зміна позначення звуку на письмі	ci→ć, ć→ci, i→j, ii→i, –(-j-), ni→ń, ń→ni, ś→si, zi→ź, ź→zi
8	зміна основи слова	–

Таблиця 3 – Фрагмент таблиці ланцюжків елементарних перетворень для іменників

NounID	GM	Chain
356	SgG	T002T006
356	SgD	T002T024
356	SgI	T002T012
356	SgL	T002T024

6 ОБГОВОРЕННЯ

Отримані ланцюжки перетворень при утворенні ГФ іменників показують крок за кроком, як саме відбувається процес словозміни. Таке подання процесу словозміни є більш наочним і зрозумілим, ніж таблиці ГФ, що містяться в електронних граматичних словниках, і може бути використане при створенні електронних засобів навчання граматики.

По-перше, на основі наведеної моделі може бути створено довідник зі словозміни, де для кожної ГФ відображатиметься відповідний ланцюжок перетворень із необхідними коментарями.

По-друге, з таблиці ланцюжків елементарних перетворень через SQL-запити можна проводити вибірку саме тих слів, у яких відбувається конкретне чергування звуків, випадання голосних, додання певного закінчення тощо. При створенні навчального курсу це дасть змогу автоматично генерувати приклади, вправи і тестові завдання відповідно до теми заняття. Крім того, за умови використання в системі достатньо великого словника, розміченого за сферами використання лексики, слова можна буде підбирати залежно від потреб і зацікавлень конкретного студента. Це забезпечуватиме індивідуалізацію процесу навчання.

Вибірка всіх слів, до яких застосовуються певні перетворення, може бути корисною не лише при наповненні електронних курсів, а й при підготовці викладачами звичайних (паперових) контрольних завдань і прикладів, а також для подальшого аналізу процесів, що відбуваються в мові, фахівцями з філології.

ВИСНОВКИ

Описаний підхід до формалізації словозміни польської мови, на відміну від існуючих, дозволяє не лише синтезувати ГФ слів, а й пояснювати цей процес крок за кроком за рахунок його декомпозиції до найпростіших елементів. Це уможливлює створення на основі поданої моделі електронних засобів навчання граматики, що надаватимуть системні знання про словозміну. Предметом подальшого дослідження є моделювання словозміни інших частин мови, аналіз і опис закономірностей застосування окремих перетворень, а також створення електронного засобу навчання на основі розробленої моделі.

ПОДЯКИ

Роботу виконано в рамках наукових досліджень кафедри інформаційних систем Національного університету харчових технологій на тему «Нові інформаційні технології в освіті».

СПИСОК ЛІТЕРАТУРИ

1. Ranta A. How Predictable is Finnish Morphology: An Experiment in Lexicon Construction : CLT Seminar, 25 September 2008 [Electronic resource] / Aarne Ranta, 2008. – 64 p. – Access mode: www.cse.chalmers.se/~aarne/talks/finnish-2008.pdf.
2. Широков В. А. Елементи лексикографії : моногр / В. А. Широков ; Укр. мов. інформ. фонд НАН України. – К. : Довіра, 2005. – 304 с.
3. Woliński M. A Relational Model of Polish Inflection in Grammatical Dictionary of Polish / M. Woliński // Human Language Technology: Challenges of the Information Society. – Berlin, Heidelberg : Springer-Verlag, 2009. – P. 96–106.
4. Słownik gramatyczny języka polskiego / [Z. Saloni, W. Gruszczycski, M. Woliński, R.Wołosz]. – Warszawa : Wiedza Powszechna, 2008. – 180 p.
5. PoliMorf – otwarty siownik morfologiczny : prezentacja [Electronic resource] / M. Wolicski, M. Miikowski, M. Ogrodniczuk [et al]. – Warszawa : IPI PAN, 2011. – 44 p. – Access mode : http://nlp.ipipan.waw.pl/NLP-SEMINAR/111205.pdf.
6. Morfologik v. 2.0 [Electronic resource]. – 2013. – Access mode : http://sourceforge.net/projects/morfologik/files/morfologik/2.0.
7. Kazojć J. Siownik frekwencyjny leksemyw V.06.2009 [Electronic resource]. – 2009. – Access mode : http://www.slowniki.org.pl/slownik-frekwencyjny-leksemow.pdf.
8. Василевская Д. Учебник польского языка / Данута Василевская, Станислав Кароляк. – СПб. : Лань, 2001. – 576 с. – (Учебники для вузов. Специальная литература).
9. Практический курс польского языка. Базовый учебник / [Я. А. Кротовская, Л. Г. Кашкуревич, Г. М. Лесная, Н. В. Селиванова]. – 2-е изд., перераб. и доп. – М. : АСТ, 2005. – 559 с.

Стаття надійшла до редакції 23.10.2015.
Після доробки 30.10.2015.

Костиков Н. П.
Ассистент кафедры информационных систем Национального университета пищевых технологий, Киев, Украина

ФОРМАЛЬНАЯ МОДЕЛЬ СЛОВОИЗМЕНЕНИЯ СУЩЕСТВИТЕЛЬНЫХ ПОЛЬСКОГО ЯЗЫКА

При создании электронных средств обучения актуальным вопросом является формализация знаний для их наглядного представления и упрощения их обработки. Формальное описание грамматики, в частности словоизменения, существующее в современных

средствах обработки естественного языка, только фиксирует грамматические формы слов, однако не анализирует процесса их образования, что необходимо для целей изучения языка.

В статье предложен подход к формализации словоизменения польского языка (на примере существительных) путем выделения простейших преобразований в процессе словоизменения. Проведен отбор слов согласно с частотным словарем польского языка. Сформирована эталонная таблица грамматических форм для выделенного набора слов. Описаны отдельные преобразования при словоизменении существительных, и с их помощью проведено моделирование процесса образования словоформ из эталонной таблицы. Полученные в результате моделирования цепочки преобразований для каждого слова сохранены в базе данных.

Полученные цепочки позволяют шаг за шагом описать образование грамматических форм слов. За счет этого обеспечивается наглядность представления процесса словоизменения для студентов, изучающих язык. Кроме того, возможен автоматический подбор слов, в которых происходит то или иное грамматическое явление, для формирования учебных упражнений, примеров и тестовых заданий.

Ключевые слова: грамматика, компьютерная лингвистика, модель языка, морфология, обработка естественного языка, электронные средства обучения.

Kostikov M. P.
Teaching assistant, Information Systems department, National University of Food Technologies, Kyiv, Ukraine

A FORMAL MODEL OF POLISH NOUNS INFLECTION

An urgent problem when creating e-learning software is knowledge formalization for its further processing and visual presentation. A formal description of grammar, in particular inflection, presented in modern means of natural language processing, only enumerates the existing grammatical forms without analyzing the process of their production, which is important for the purposes of language learning.

The approach to Polish inflection formalization (by the example of nouns) which consists in separating out the basic elements of inflection process is proposed in the presented paper. The selection of words according to the frequency dictionary of Polish is performed. The standard table of grammatical forms for the selected words is formed. Individual transformations in the process of nouns inflection are described. With the help of the described transformations, the modeling of word forms generation process is performed. The resulting chains of transformations for each word are saved into a database.

The obtained chains allow of describing the process of inflection step-by-step. It makes the knowledge presentation to the language students more clear and visual. Besides, the automatic selection of words which go through certain transformations is possible for the purposes of generating learning exercises, examples, and test questions.

Keywords: computational linguistics, computer-assisted language learning, grammar, language model, morphology, natural language processing.

REFERENCES

1. Ranta A. How Predictable is Finnish Morphology: An Experiment in Lexicon Construction : CLT Seminar, 25 September 2008 [Electronic resource], 2008, 64 p. Access mode : www.cse.chalmers.se/~aarne/talks/finnish-2008.pdf.
2. Shyrokov V. A. Elementy leksykohrafii : monohr. Ukrainian Lingua-Information Fund, NAS of Ukraine. Kiev, Dovira, 2005, 304 p.
3. Woliński M. A Relational Model of Polish Inflection in Grammatical Dictionary of Polish, *Human Language Technology: Challenges of the Information Society*. Berlin, Heidelberg, Springer-Verlag, 2009, pp. 96–106.
4. Zygmunt Saloni, Włodzimierz Gruszczyński, Marcin Woliński, Robert Wołosz Siownik gramatyczny jкzyka polskiego. Warszawa, Wiedza Powszechna, 2008, 180 p.
5. Marcin Woliński, Marcin Miłkowski, Maciej Ogrodniczuk [et al.]. PoliMorf – otwarty siownik morfologiczny : prezentacja [Electronic resource]. Warszawa, IPI PAN, 2011, 44 p. Access mode : http://nlp.ipipan.waw.pl/NLP-SEMINAR/111205.pdf.
6. Morfologik v. 2.0 [Electronic resource], 2013, Access mode : http://sourceforge.net/projects/morfologik/files/morfologik/2.0.
7. Kazojć J. Słownik frekwencyjny leksemów V.06.2009 [Electronic resource], 2009, Access mode : http://www.slowniki.org.pl/slownik-frekwencyjny-leksemow.pdf.
8. Danuta Vasilevskaya, Stanislav Karolyak Uchebnik pol'skogo yazyka. Sankt-Peterburg, Lan', 2001, 576 p.
9. Krotovskaya Ya. A., Kashkurevich L. G., Lesnaya G. M., Selivanova N. V. Prakticheskij kurs pol'skogo yazyka. Bazovyj uchebnik. 2nd ed. Moscow, AST, 2005, 559 p.