# КОМП'ЮТЕРНА ЛІНГВІСТИКА

## ПОБУДОВА ПАРАДИГМАТИЧНОГО СЛОВНИКА ДЛЯ СИНТЕЗУ СЛОВОФОРМ СУЧАСНОЇ УКРАЇНСЬКОЇ МОВИ

© **Ольга Шипнівська, 2010**
к. філол. н., Інститут філології Київського національного університету
імені Тараса Шевченка (Київ)
УДК 811.161.2.373′322.2

У статті запропоновано підхід до створення парадигматичного словника для синтезу словоформ сучасної української мови на основі словника квазізакінчень. Розглянуто наявні методи синтезу та специфіка застосованої методики генерації текстових словоформ.
**Ключові слова:** парадигматичний словник, словник квазізакінчень, словник словоформ, процедура синтезу словоформ.

In the article the attitude for making of the paradigmatic dictionary for generation of Ukrainian word forms on the base of the kwasiflexes dictionary is proposed. Available methods of synthesis and the peculiarity of our method of Ukrainian word forms generation are described.
**Keywords:** the paradigmatic dictionary, the dictionary of kwasiflexes, the dictionary of word forms, the procedure of word forms' synthesis.

Важливим компонентом лінгвістичного процесора є синтетичні словники морфологічного рівня мовної системи, зокрема парадигматичний. Розроблення ефективної системи машинного перекладу (СМП) вимагає удосконалення цього етапу опрацювання мовної інформації. У статті пропонуємо процедуру створення парадигматичного словника для синтезу словоформ сучасної української мови відповідно до принципів розроблення знання-орієнтованої системи машинного перекладу.

Проблема генерування природномовного тексту повсякчас примушувала дослідників шукати та застосовувати найприйнятніші шляхи синтезу текстових словоформ [1; 2]. На сьогодні відомо три

основні підходи щодо морфологічного синтезу: словниковий метод, алгоритмічний метод та метод морфологічного синтезу на базі словника квазізакінчень [9].

Словниковий метод базується на використанні словника словоформ із відповідною морфологічною інформацією. Такий підхід застосовують здебільшого в системах з невеликим обсягом словника. У переважній більшості СМП залучають алгоритмічний метод синтезу словоформ. У цьому разі для програми генерування головним компонентом є парадигматичний словник. Словникова стаття такої лексикографічної системи містить усю парадигму заданої основи.

Загалом процедуру синтезу словоформ розглядають як результат роботи модуля автоматичного морфологічного аналізу (АМА). Відповідно, обрана дослідниками комп’ютерна модель аналізу тексту визначає й специфіку програми синтезу. Для української мови системи генерування з урахуванням алгоритмічного підходу розробляли в межах створення процедур автоматичного морфологічного аналізу в Інституті мовознавства ім. О. О. Потебні НАН України для опрацювання наукових текстів російською та українською мовами [4; 9; 10] та в Українському мовно-інформаційному фонді НАН України, де залучено понад 3 тис. парадигматичних класів [7]. Відома модель синтезу парадигм українського дієслова, яку запропонував С. Лук’янчук [8].

Як і для системи АМА, оцінка ефективності процедури синтезу текстових словоформ залежить від обсягів словника, несуперечливості інформації, швидкості опрацювання текстів і можливості аналізу нових слів [1; 2; 3]. Крім того, знання-орієнтований підхід до розроблення СМП вимагає побудови відкритої системи автоматичного синтезу завдяки використанню наявних в текстах лінгвістичних знань та знань з предметної галузі [5].

У розроблюваній нами системі автоматичного морфологічного аналізу використано флективний підхід за словником квазізакінчень, отриманого на базі словника словоформ. Словник словоформ нараховує на сьогодні понад 32 тис. одиниць[1].

[1] Детально про АМА див. [6].

Його джерельною базою стали тексти військово-спеціальних та військово-гуманітарних наук різних жанрів та матеріали Internet[1].

Розроблений метод подання граматичної інформації в словниковій статті базується на позиційно-цифровому кодуванні. У такий спосіб кожна аналізована словоформа отримує свій код, який містить інформацію про її частиномовну належність та конкретне граматичне значення.

Класифікація лексико-граматичних класів, застосована нами у системі АМА, зорієнтована на те, що результати слугують вихідними даними для автоматичного синтаксичного, лексичного та семантичного аналізів. Звісно, виокремлюємо традиційні частини мови (іменник, дієслово й т. ін.). Зміни, які ми внесли, стосуються класів:

- дієслова – в окрему групу виділяємо дієслова минулого часу, неминулого часу, інфінітив, вказуючи на особливості керування, дієслова наказового способу; в окремі групи виділено дієслівні форми – дієприкметник та дієприслівник;
- числівника (виокремлено числівник-іменник, числівник-прикметник);
- для займенників ураховано як характер їхніх значень, так і специфіку словозміни та функціонування в мові (розглядаємо займенники, що мають прикметниковий тип відмінювання (деякий, кожний, всякий), займенники-іменники (вона, він), особові займенники);
- для прийменника додано лексико-граматичну категорію керування;
- виокремлено прислівник у порівняльній формі.

У таблиці 1 подано лексико-граматичні класи та їх граматичні категорії, обраної нами класифікації[2].

[1] Розробка словника здійснюється в рамках наукової теми зі створення знання-орієнтованої системи машинного перекладу текстів військової тематики.

[2] Система АМА розробляється для української, російської, англійської мов, а тому взято такі класи: артикль, герундій.

Таблиця 1. Коди лексико-граматичних класів та їх граматичні категорії

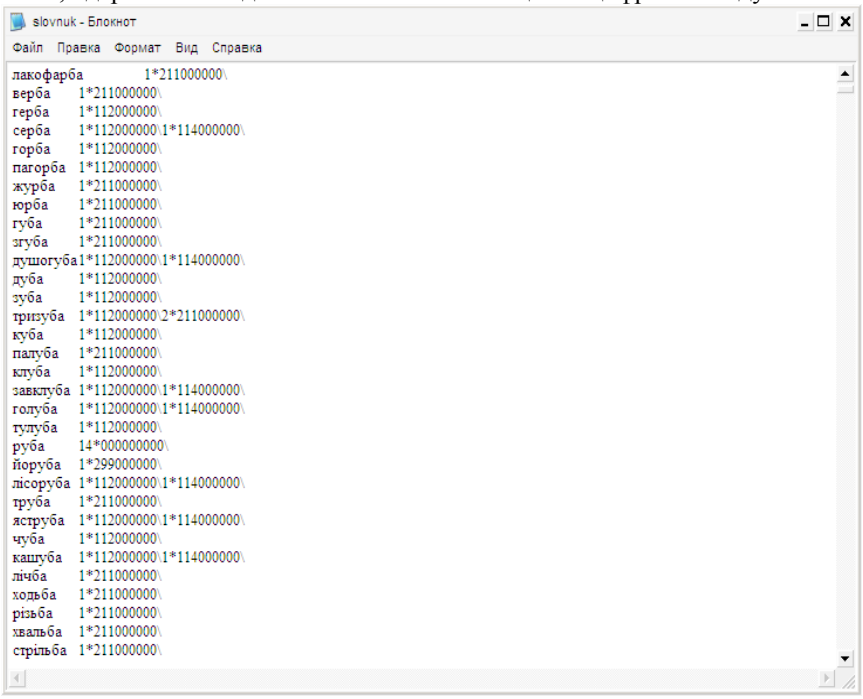
| Код л-г класу | Лексико-граматичний клас | **ГРАМАТИЧНІ КАТЕГОРІЇ** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Рід | Число | Відмінок | Особа | Вид | Стан | Ступінь | Час | Зворотність | |
| 1 | *Іменник* | + | + | + | | | | | | | |
| 2 | *Прикметник* | + | + | + | | | | | | | |
| 3 | *Числівник-іменник* | + | | + | | | | | | | |
| 4 | *Числівник-прикметник* | + | + | + | | | | | | | |
| 5 | Займенник особовий | + | + | + | + | | | | | | |
| 6 | Займенник-іменник | | | + | | | | | | | |
| 7 | Займенник-прикметник | + | + | + | | | | | | | |
| 8 | Дієслово минулого часу | + | + | | | + | | | | + | |
| 9 | Дієсл. неминулого часу | | + | | + | + | + | | + | + | |
| 10 | Інфінітив | | | | | + | | | | | |
| 11 | Дієслово наказ. способу | | + | | | | | | | | |
| 12 | Дієприкметник | + | + | + | | | + | | + | | |
| 13 | Дієприслівник | | | | | | | | | + | |
| 14 | Прислівник | | | | | | | | | | |
| 15 | Порівняльний прикметник | | | | | | | + | | | |
| 16 | Присвійний прикметник | + | + | + | | | | | | | |
| 17 | Короткий прикметник | + | + | | | | | | | | |
| 18 | Дієприкметник на -но, -то | + | + | | | + | | | + | | |
| 19 | Прислівник у порівн. формі | | | | | | | + | | | |
| 20 | Частка | | | | | | | | | | |
| 21 | *Модальне слово* | | | | | | | | | | |
| 22 | Артикль | | | | | | | | | | + |
| 23 | Прийменник | | | + | | | | | | | |
| 24 | Сполучник | | | | | | | | | | |
| 25 | Герундій | | | | | | | | | | |

Порівняймо, у класифікації, запропонованій в монографії “Морфологічний аналіз наукового тексту на ЕОМ” (К.: Наук. думка,

1989), не виділяються в окремі групи “дієслово минулого/неминулого часу”, “інфінітив”, “прислівник у порівняльній формі”, “порівняльний прикметник”, дієприкметник розглядається у групі “ад’єктив”, аналогічно до нашого підходу кваліфікується прийменник – з урахуванням специфіки керування, дієприслівник включено до парадигми дієслова.

На малюнку 1 подано фрагмент словника словоформ української мови із зазначеною граматичною параметризацією.

Малюнок 1. Фрагмент словника словоформ сучасної української мови, одержаного за допомогою АМА з позиційно-цифровим кодуванням.




Дослідники встановили, що в мовах флективного типу з розгалуженою системою словозміни 95 % морфологічної інформації про словоформу зосереджено в кінці слова, в його квазізакінченні. Зважаючи на це, автоматичний морфологічний аналіз саме на базі словника квазізакінчень є найпрогресивнішим, його застосовують у

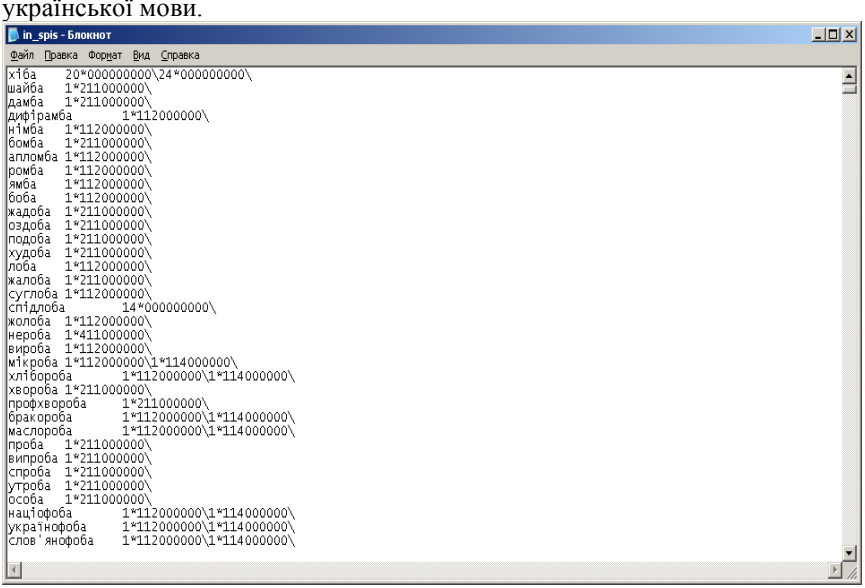
сучасних промислових системах автоматичного аналізу текстової інформації. Спосіб подання граматичної інформації не за повним словом, а за кінцевою послідовністю літер, тобто застосований нами принцип флективного аналізу, дозволяє значно скоротити обсяг аналітичного словника. Такий словник здатен розпізнавати з тією ж самою точністю, що і на основі словника словоформ, всі лексичні одиниці вхідної мови.

Словник квазізакінчень формувався на основі інверсного словника словоформ. На малюнку 2 подано фрагмент цього словника. Словникова стаття словника квазізакінчень має такий формат подання:

$$<\{F\}_n^1 \quad K_l * \{K_g\}>,$$

де $\{F\}_n^1$ – ланцюг літер вхідного слова, розташований в інверсному порядку, $n$ – остання літера в слові, довжина ланцюга загалом може бути рівною довжині вхідної словоформи. $K_l$ – код лексико-граматичного класу вхідної словоформи, $K_g$ – множина кодів значень граматичних категорій, які визначаються для $K_l$.

Малюнок 2. Фрагмент інверсного словника словоформ сучасної української мови.



За інверсним словником словоформ було побудовано лематизаційний словник, за яким кожна текстова словоформа отримувала квазізакінчення канонічної, вихідної форми. На малюнку 3 подано фрагмент лематизаційного словника.

Малюнок 3.Фрагмент лематизаційного словника сучасної української мови.

**аб 1*21/ба/1*11/б/2*21/бий/ба/бе/1*41/ба/**
**аберго 1*11/огріб/**
**абер 1*11/реб/**
**або 1*11/іб/1*41/оба/1*21/оба/**
**абол 1*11/лоб/1*21/лоба/**
**абор 1*41/роба/1*11/ріб/**
**аборк 1*11/кроб/**
**аборо 1*11/ороб/1*21/ороба/**
**абоф 1*11/фоб/**

Парадигматичний словник для генерування текстових словоформ сучасної української мови був створений на основі лематизаційного словника як його пряме віддзеркалення. Словникова стаття парадигматичного словника має такий вигляд:

$$<\{F_i\} \quad K_l * K_g(P)>,$$

де $\{F\}_n^l$ – ланцюг літер вхідного слова, розташований в інверсному порядку, $K_l$ – код лексико-граматичного класу вхідної словоформи, $K_g(P)$ – множина квазізакінчень граматичних форм, які визначаються для $K_l$. Так, для лексеми *імпреза* словникова стаття має вигляд:

*аз 1*21/за/зи/зі/зу/зою/зі/зо/зи/з/зам/зами/зах/.*

Така словникова стаття не залежить від конкретної основи й описує парадигму певного класу основ (слова: гіпотеза, теза, криза, віза, валіза та ін.), що мають однакову словозмінну парадигму. У разі паралельних форм через знак "//" подано відповідний варіант, як, наприклад, для словникової статті:

*к*
*1*11/к/ка/ку//кові/к//ка/ком/ку/кові/ки/ків/кам/ки//ків/ками/ках/*,

за якою можуть бути синтезовані лексеми *літак* та *полковник*. Представлена словникова стаття об’єднує слова, різнорідні за семантикою: істоти та неістоти. Відповідно, під час синтезу із залученням додаткової інформації генеруються форми родового та знахідного відмінків. Форми ж давального та місцевого визначаються як паралельні.

Звісно, на основі графемного аналізу було враховано всі морфонологічні зміни, властиві основам змінюваних слів певних частин мови. Так, за попередньою словниковою статтею, попри наявність кінцевої графеми “к”, не можна генерувати лексеми *садок, візок, валок*, оскільки в їхній основі є випадний “о”. Усі ці лексеми будуть породжені за формальними правилами генерування текстових слів (відповідно):

*код*
*1*11/док/дка//ку/ку//кові/док//дка/ком/ку//кові/ки/ків/кам/ки//ків/ками/ках/;*

*коз*
*1*11/зок/зка//зку/зку//зкові/зок//зка/зком/зку//зкові/зки/зків/зкам/зки//зків/зками/зках/;*

*кол*
*1*11/лок/лка//лку/лку//лкові/лок//лка/лком/лку//лкові/лки/лків/лкам/лки//лків/лками/лках/.*

Для багатьох випадків квазізакінчення цілком збігається зі словом, напр., іменники чоловічого роду на “р”, зокрема іменники мішаної та м’якої групи, які різняться від твердої групи відмінювання: *бджоляр*, *бетоняр*; дієслова: *дати, стати*.

Об’єктом синтезу морфологічного рівня є будь-яка словоформа. Розроблюваний нами словник зорієнтований і на породження дієслівних форм: дієприкметника та дієприслівника. Так, словникова стаття слова *вести* містить, крім інших, ще й ці форми дієслова:

*итсев*
*/веди/ведімо/ведіть/веду/ведеш/веде/ведемо/ведете/ведуть/вестиму/вестимеш/вестиме/вестимемо/вестимете/вестимуть/вів/вела/вело/вели/вівши/ведений/*.

Алгоритм синтезу працює так:

1. В аналізованого слова за словником виділяється квазізакінчення.
2. Визначене квазізакінчення відсікається.
3. Визначається граматична форма, яку потрібно генерувати.
4. Підраховується номер відповідної граматичної форми.
5. Вибирається необхідне квазізакінчення із словника.
6. До слова приписується встановлене квазізакінчення.

Переваги обраної нами процедури автоматичного синтезу словоформ вбачаємо у швидкості та простоті алгоритму, що значною мірою зумовлюється обраним позиційно-цифровим кодуванням. Можливість опрацювання нових, невідомих слів робить систему відкритою та значно поліпшує ефективність програми машинного перекладу. Варто зауважити, що ще й досі триває робота з побудови оптимального списку квазізакінчень на основі нової вибірки текстів різної тематики.

## Література

1. Апресян Ю. Д., Кулагина О. С. Проблемы разработки систем МП // Актуальные вопросы практической реализации систем автоматического перевода. – М.: Изд-во Моск. ун-та, 1982. – С. 5–24.
2. Апресян Ю. Д., Богуславский И. М. и др. Лингвистическое обеспечение в системе автоматического перевода третьего поколения. – М., 1978. – 74 с.
3. Гельбух А. Ф., Сидоров Г. О. К вопросу об автоматическом морфологическом анализе флективных языков // Ел. режим доступу: www.dialog-21.ru/Archive/2005
4. Грязнухіна Т. О., Нікула М. В. Система автоматичного морфологічного аналізу українського наукового тексту // Проблеми українізації комп'ютерів. Матеріали 2-ї міжнародної конференції. – Київ, 1993. – С. 42–46.
5. Замаруєва І. В. Комп'ютерна модель розуміння природномовної текстової інформації // Проблемы программирования. –1999. – №2. – С. 96–102.
6. Замаруєва І. В., Шипнівська О. О. Морфемна обробка текстів в системах машинного перекладу // Вісник Київського національного університету імені Тараса Шевченка. Військово-спеціальні науки. – К., 2008. – №20. – С.61–63.

7. Корпусна лінгвістика: Монографія / Широков В. А., Бугаков О. В., Грязнухіна Т. О., Костишин О. М., Кригін М. Ю., Любченко Т. П., Рабулець О. Г., Сидоренко О. О., Сидорчук Н. М., Шевченко І. В., Шипнівська О. О., Якименко К. М.; Український мовно-інформаційний фонд НАН України. – К.: Довіра, 2005. – 471 с.
8. Лук’янчук С. Комп’ютерна модель парадигматичних класів дієслів // Українське мовознавство. – 2000. – Вип. 22. – С. 82–85.
9. Олексієнко Л., Дарчук Н. Лематизація парадигм іменників української мови // Проблеми українізації комп’ютерів. Матеріали 2-ї міжнародної конференції. – Київ, 1993. – С. 42–46.
10. Перебийнос В. И. Сведение парадигм в размеченном и неразмеченном тексте // Морфологический анализ научного текста на ЭВМ. – К.: Наук. думка, 1989. – С. 20–250.
11. Хайрова Н. Ф., Замаруева И. В. Машинный перевод. – Харьков: Око, 1998. – 80 с.