УДК 81.′322.2

АВТОМАТИЗАЦІЯ СКЛАДАННЯ ГРАМАТИЧНИХ СЛОВНИКІВ ЛЕКСИКИ ТЕМАТИЧНИХ ТЕКСТІВ

Н.М. Міщенко, О.Д. Феліжанко, Н.М. Щоголева

Інститут кібернетики імені В.М. Глушкова НАН України,
03680, Київ, проспект Академіка Глушкова, 40,
тел. (38)(044)526 2278, e-mail: nat@d100.icyb.kiev.ua

Пропонуються програмні засоби для складання граматичних словників лексики тематичних текстів за результатами безсловникового морфологічного аналізу таких текстів. Розглядаються особливості структури та функції запропонованих програмних засобів.

In the article the software components for generation of grammatical dictionary of professional texts lexemes using the results of morphological analysis for professional texts without dictionary are proposed. The structure and functions peculiarities of software components proposed are discussed.

Вступ

Граматичні словники використовуються в програмах перекладу та для лінгвістичних досліджень текстів. Кожна стаття граматичного словника, складання якого є метою нашого дослідження, – це парадигма лексеми (усі словоформи, породжувані відмінюванням чи дієвідмінюванням), а також морфологічна класифікація лексеми. Парадигма представляється у словнику у вигляді основи лексеми, з якою асоціюється кортеж закінчень усіх похідних словоформ, та мнемонічна назва морфологічних характеристик лексеми.

Складання граматичних словників виконується системою програм FEST, яка є морфологічною підсистемою лінгвістичної системи LIS [1], створеної в Інституті кібернетики НАН України. До складу системи FEST, крім основних програм, а саме, морфологічного аналізу (МА) та генерації формального опису статей граматичного словника, входять кілька допоміжних програм, призначених для інформаційного забезпечення основних.

Традиційно програми МА орієнтуються на аналіз текстів певною флективною мовою, послуговуються відповідними морфологічними таблицями та словником лексики. Програма МА MORF системи FEST відрізняється від традиційних тим, що її програмна частина універсальна, тобто придатна для аналізу текстів кількома флективними мовами. Загальною для всіх мов є також структура морфологічних таблиць та словників. Настроюванням універсальної частини на аналіз текстів конкретною мовою відбувається заповненням табличних структур морфологічною інформацією конкретної мови та побудовою її словника.

Користувачеві програми MORF не потрібно знати структуру комп'ютерних морфологічних таблиць та словників, оскільки їх заповнення автоматизовано.

Переходимо до короткого опису інформації та функцій окремих програм системи FEST, які беруть участь у формуванні граматичного словника.

Морфологічні таблиці

Для автоматичної побудови морфологічних таблиць створена метамова специфікації морфологічної інформації лексем. Специфікація морфології morf.grm містить такі розділи:

– алфавіт мови (малі та великі літери);

– скорочені мнемонічні назви відмінків для іменних частин мови, осіб для дієслів та їхні коди;

– мнемонічні назви кортежів закінчень – ланцюжки символів, якими позначені морфологічні ознаки класу лексем, що приймають закінчення кортежу, та їхні коди;

– власне кортежі, що починаються назвами кортежів.

Специфікація morf.grm програмою GENm перетворюється на морфологічні таблиці morf.tbl.

Наразі використовуються специфікації морфології української та російської мов (файли, відповідно, umorf.grm і rmorg.grm).

Далі наводимо фрагменти розділів специфікації морфології української мови umorf.grm, розділені рядками крапок. Нуль на початку рядків означає, що у рядку коментар. Нуль на початку алфавіта використовується для позначення відсутності закінчення (“нульове” закінчення) у словоформах змінюваних лексем. В інших рядках коментарі знаходяться між парами символів /* та */.

```
0    file project/morf/test/umorf.grm 19/12/05
0
196  0абвгдеєжзиіїйклмнопрстуфхцчшщюяь';
197  0АБВГДЕЄЖЗИІЇЙКЛМНОПРСТУФХЦЧШЩЮЯЬ';
0
```

```
0    Коди    Назви відмінків та осіб
198     1    он;   /* називний відм. одн. */
198     2    ор;   /* родовий відм. одн. */
. . . . . .
198     1    він;     /* чол. рід, 3-я особа одн. */
198     2    вона;  /* жін. рід, 3-я особа одн. */
. . . . . .
 0     Коди   Назви кортежів
199   1211   іж1т_1;    /* ім., жін. р.,1 відм., тв. гр., (мов-а) */
199   1212   іж1тж_2;  /* ім. жін. р.,1 відм., тв. гр., жив. (людин-а) */
199   1213   іж1ш_3;   /* ім. жін.р.,1 відм., міш. гр. (задач-а) */
199   1214   іж1м_4;   /* ім. жін.р.,1 відм.,  м'яка гр. (таблиц-я) */
. . . . . .
0      Кортежі закінчень
4 іж1т_1      а  и  і  у  .  ою  і  и  0   ам  и  .  ами  ах   (мова);
5 іж1тж_2   а  и  і  .  у  ою  і  и  ей  ям  .  ей  ьми  ях  (людина);
6 іж1ш_3    а  і  і  у  .  ею  і  і  0   ам  і  .  ами  ах   (задача);
7 іж1м_4    я  і  і  ю  .  ею  і  і  ь  ям  і  .  ями  ях  (таблиця);
```

На основі кортежів закінчень програма GENm будує також таблицю OMON, у якій кожне закінчення супроводжується списком назв кортежів, де воно зустрічається.

Словник основ

Морфологічний аналіз текстів виконується за словником основ, який формується автоматично на основі специфікації основ lex.spc, кожна з яких в словнику супроводжується граматичною інформацією, що дозволяє розпізнавати словоформи з цією основою.Такою інформацією є мнемонічна назва кортежа закінчень, що їх приймає основа, суфікси з назвою відмінка, у якому він може бути у словоформі. Якщо слово у різних формах має різні суфікси, то досить указати відмінки чи особи для всіх суфіксів, крім останнього, який означає, що останній вживається у решті відмінків.

Якщо всі словоформи деякої лексеми мають один і той же суфікс, то він додається до основи. Слід зауважити, що в системі FESTдо суфіксів відносять також змінну частину основи, що межує з закінченням. Наприклад, лексема "наслідок" має основу "наслід", суфікси "ок", "к" і "нульове" закінчення. Специфікація цієї основи поданадалі. Наведемо приклади специфікації кількох основ:

```
мов => * : іж1т_1 ;
модел => * : іж3м_ь  "л" (оо) ; /* в орудному відмінку – подвійне 'л' */
наслід => * : іч2_0 "ок" (он, оз) "к" .
```

Специфікації основ мови складаються вручну носіями цієї мови. Передбачене також автоматизоване формування специфікацій основ програмою GENspc.

Перевірка правильності специфікації виконується за допомогою програми GENw – генератора словоформ. Отриманий результат аналізується носієм мови, неправильні словоформи свідчать про помилку в специфікації відповідної основи.

Програма GENlex перетворює специфікації основ на статті граматичного словника D.

Морфологічний аналіз може виконуватися і без словника або при неповному словнику, але з морфологічними таблицями. Оскільки лінгвістична система FEST призначена для лінгвістичних досліджень тематичних текстів, лексика яких та її граматичні характеристики можуть бути відсутніми в існуючих словниках, то для використання цієї системи доводиться граматичні словники створювати. Така робота виконується за участю самої системи FEST. Для цього програма MORF уподовж МА формує список відсутніх у словнику основ словоформ, відкидаючи закінчення з кінця словоформ. Список основ невідомих словоформ є інформацією для автоматичного формування специфікацій таких основ, тобто знаходження мнемонічних назв відповідних кортежів, у яких закодовані необхідні морфологічні характеристики.

Слід зауважити, що загальновживані слова тематичних текстів, тобто, службові та незмінні слова зустрічаються частіше за терміни, які, як відомо, мають найвищу частоту входжень порівняно з іншою повнозначною лексикою. За відсутності службової та незмінної лексики в словнику список невідомих основ буде ускладнений невідомими словами цього типу з відділенням від деяких з них кінцевих літер, які збігаються з закінченнями (несправжня омонімія закінчень). Специфікувати незмінну і службову лексику та автоматично ввести її в словник нескладно, до того ж це робиться один раз для текстів різної тематики. Отже, будемо вважати, що після морфологічного аналізу отримуємо список основ та закінчень лише відсутньої в словнику повнозначної лексики тексту, що аналізується.

Наводимо початкову частину файла з незнайденими програмою MORF основами словоформ у тексті abdw.txt (53 Кб) з морфологічними таблицями umorf.tbl та словником, побудованим за специфікацією службових та незмінних слів uw-aux.spc:

```
"abdw.txt, umorf.tbl, uw-aux.spc";
      2   5   0    3       оцінк;
@     2   5   0    3       якісн;
      3   9   0    1       ГЕНЕРАТОР;
@     3   4   0   77       МОВН;
@     3   8   0   65       ПРОЦЕСОР;
      3   8   0   77       ФЛЕКТИВН;
@     3   2   0    5       МО;
      5   8   0  418       розгляда;
@     5   5   0   11       засоб;
@     5  10   0   80       автоматичн;
@     5   8   0   15       генераці;
@     5   4   0   77       мовн;
. . . . . .
```

Подаємо деякі пояснення. У першому стовпчику відсутність символа свідчить про те, що відповідна основа відсутня в словнику. У цьому ж стовпчику безпосередньо за рядком без символа може бути символ @ або кілька таких символів один за другим. Так позначаються незнайдені основи сусідніх (в тексті) словоформ, перша з яких починається у рядку без символа @ у першому стовпчику і закінчується у рядку з останнім символом @ перед наступним пустим. Так, у нашому фрагменті маємо, наприклад, такі основи сусідніх словоформ: "оцінк якісн", "генератор мовн процесор". Другий стовпчик – це номер рядка у тексті abdw.txt. Третій стовпчик – число літер в основі. Четвертий – нуль для невідомих основ, для відомих – початок суфікса (якщо такий є в словоформі) в масиві суфіксів. П'ятий – початок закінчення у масиві закінчень, шостий містить основи незнайдених словоформ.

Визначення морфологічних характеристик.

Для визначення характеристик лексем за списком незнайдених основ з закінченнями програмою FRlist формується частотний список невідомих основ у порядку зменшення їх вживання в тексті. Подаємо початковий фрагмент частотного списку, отриманого після морфологічного аналізу тексту abdw.txt за морфологічними таблицями umorf.tbl та словником незмінних слів uw-aux.spc.

```
1)   55  0.781%   153  словоформ (-0, -и, -ами, -ах, -ам, -ою, -а)
2)   47  0.667%   127  переклад (-ах, -у, -0, -ів, -ом, -и, -ати, -і)
3)   43  0.610%   141  словник (-ів, -а, -и, -ові, -у, -ом, -0)
4)   35  0.497%    19  мов (-и, -а, -ою, -ами, -ах, -у)
5)   34  0.483%   127  текст (-ах, -ів, -і, -ам, -у, -0, -ові, -ами)
```

У двох стовпчиках після порядкових номерів рядків подані відповідно частота вживання словоформ у числовому вимірі та відсоток частоти до загального числа слововживань у даному тексті.

У частотному списку з кожною основою асоціюється множина закінчень, з якими вживалися словоформи в тексті. Щоб визначити назву кортежу закінчень словоформ для кожної основи, слід звернутися до таблиці омонімів закінчень OMON. Нагадаємо, що ця таблиця для кожного закінчення містить список назв кортежів, у яких воно зустрічається. Отже, з кожною основою асоціюється той кортеж закінчень, назва якого входить в списки омонімів всіх її закінчень. За результатами пошуку програма GENspc формує специфікацію лексеми із знайденою назвою кортежу, яка подається в режимі on-line генератору словоформ GENw для перевірки правильності вибору.

Після обробляння таким чином усіх основ з частотного списку отримуємо специфікацію lex.spc, на основі якої програма GENlex доповнює словник системи FEST новими основами.

Якщо частота вживання словоформ певної лексеми низька, то може трапитися так, що множина закінчень містить недостатню їх кількість для виявлення назви єдиного кортежу. Якщо назв кортежів виявиться більше одного, припустимо два-три, то релевантний кортеж можна вибрати за результатом згенерованих програмою GENw словоформ однієї і тієї ж лексеми за кількома кортежами в режимі on-line (рис. 1).

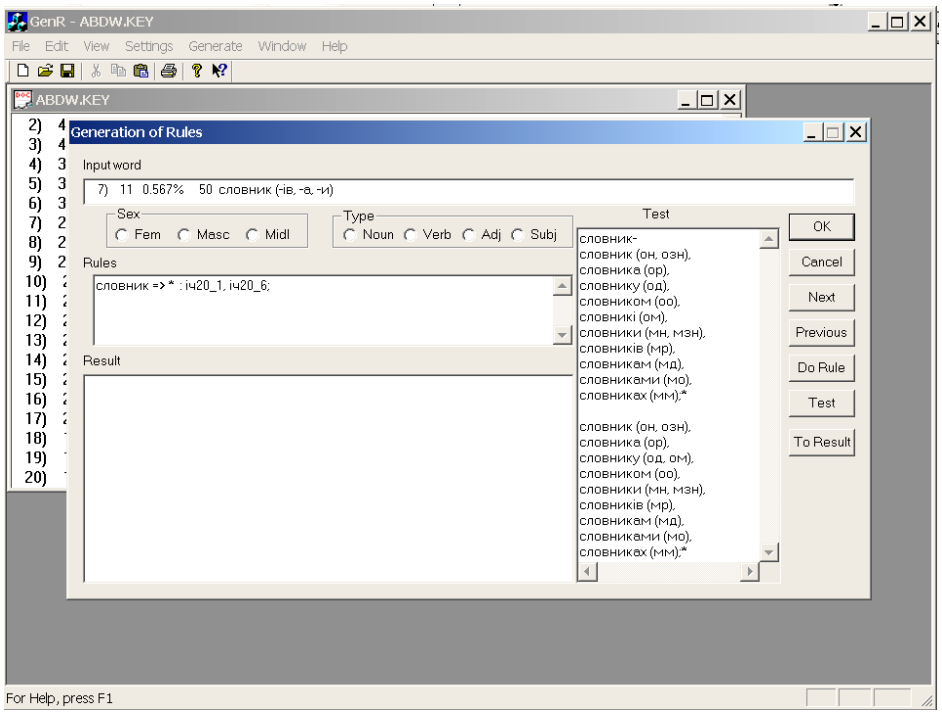

Рис. 1

Вибраний варіант показано на рис. 2.

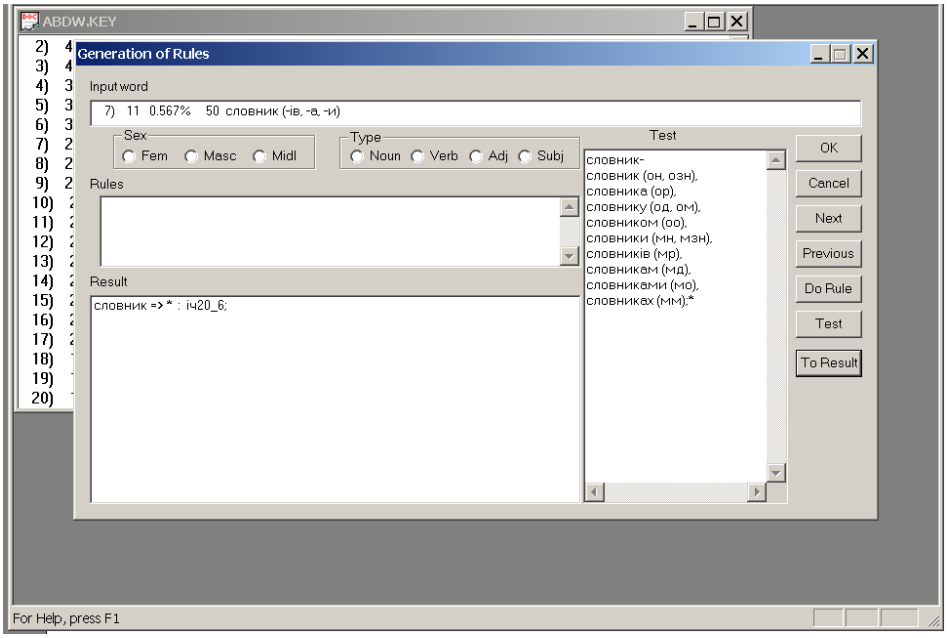

Рис. 2

Інший варіант – залишити основу незанесеною в словник до обробляння наступного тексту цієї ж тематики або ж сформувати специфікацію вручну як і тоді, коли до основ додаються суфікси не у всіх відмінках чи особах.

Наступні тексти цієї тематики подаються на аналіз з використанням уже розширеного словника. Ітеративний процес врешті решт приведе до побудови повного словника певної тематичної галузі.

На рис. 3 подаємо схему всього циклу побудови граматичного словника за умови, що морфологічні таблиці мови вже згенеровані. Подвійними лініями обведені назви програм, а одинарними – назви інформації. Стрілки показують напрям руху інформації та порядок її обробляння. Робота починається з подання на вхід програмі MORF тематичного тексту. Якщо метою експерименту є поповнення словника, то керування передається програмам, що опрацьовують незнайдені слова, як описано вище.

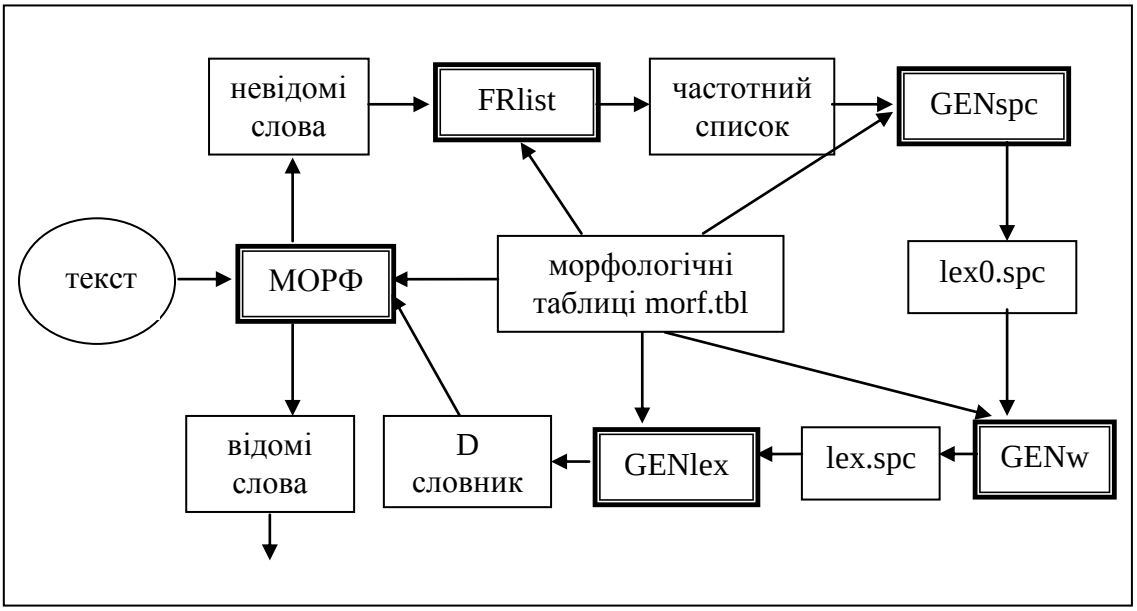

Рис. 3. Схема функціонування лінгвістичної системи LIS

Знайдені словоформи використовуються для лінгвістичного дослідження тематичних текстів за допомогою інших засобів системи LIS, а саме, для перекладів фахових текстів з російської мови на українську, для статистичних досліджень фахових текстів з метою пошуку термінів та визначення тематики текстів [2].

1. *Щоголева Н.М., Міщенко Н.М., Феліжанко О.Д.* Особливості перекладу українською наукових текстів з інженерії програмування // Проблеми програмування. Матеріали 6-ї Міжнар. конф. УкрПРОГ''2008, 27–29 травня 2008. – № 2–3. – Київ, 2008. – С. 261–269.
2. *Мищенко Н.М., Щеголева Н.Н.* О задаче семантического индексирования тематических текстов // Proc. XI-th Intern. Conf. "Knowledge-Dialog-Solution", June 20–30, 2005, Varna, Bulgaria. Volume II / FOI-Commerce, Sofia, 2005. – P. 347–350.