УДК 004.43(075.8)

А.Б. Романюк, В.М. Макар
Національний університет "Львівська політехніка",
кафедра систем автоматизованого проектування

РОЗРОБЛЕННЯ ДВОРІВНЕВОГО МОРФОЛОГІЧНОГО АНАЛІЗАТОРА УКРАЇНСЬКОЇ МОВИ У ПРОГРАМНОМУ СЕРЕДОВИЩІ PC-KIMMO

Останнім часом у лінгвістичному забезпеченні САПР спостерігається тенденція створення діалогових мов, наближених до природної мови. Використання природної мови зумовлює необхідність розроблення морфологічних, синтаксичних та семантичних аналізаторів. У статті описано розроблення морфологічного аналізатора української мови на основі системи PC-KIMMO.

In the last years in the CAD linguistic providing there is a tendency of creation of dialog languages of close to the human language. The use of human language predetermines the necessity of development of morphological, syntax and semantic analyzers. In this article development of morphological analyzer of Ukrainian is described on the basis of the system PC-KIMMO.

Вступ

Серед відомих розробок морфологічних аналізаторів найпоширеніша розроблена фінським дослідником Кіммо Коскенніємі система KIMMO, яку також називають дворівневою моделлю. На базі цієї моделі створено як інструментальні засоби, які дають змогу без додаткового програмування розробляти описи морфологічних структур природних мов, так і бібліотеки процедур, які уможливлюють використання морфологічних процедур у прикладних [1].

У системі реалізований погляд на опис морфології природної мови як на формально-математичну задачу, що дало змогу використовувати виключно математичний апарат при розробленні моделі та оцінці її обчислювальної ефективності. Морфологічний аналізатор і синтезатор у KIMMO-моделі – це перетворювачі, які працюють за принципом скінченного автомата.

До переваг KIMMO-моделі насамперед належить її високий ступінь мовної незалежності за більша, ніж в інших моделях, простота складання правил перетворень. Мовна незалежність та засновані на строгих і добре відомих програмістам математичних конструкціях механізми реалізації сприяли розробленню різних версій ефективного програмного забезпечення системи програмістами з різних країн і різними мовами програмування. Наявність ефективних програмних засобів сприяла розробленню різних версій лінгвістичного забезпечення для багатьох мов світу. Отже, на базі моделі KIMMO виник унікальний для лінгвістичної практики феномен поширеної стандартної мовонезалежної морфологічної системи із розвинутим набором програмних засобів і добре вивченим корпусом лінгвістичних описів. Тому вивчення можливостей KIMMO-моделі для створення морфологічного аналізатора української мови є актуальної науковою задачею, яка має також значну практичну цінність [5].

Експериментальні реалізації морфологічних процесорів виконані з використанням KIMMO-моделі у програмних середовищах PC-KIMMO, PY-KIMMO для японської, арабської, шведської, німецької, фінської, англійської, аккадської, румунської, російської, церковнослов'янської, французької, іспанської, івриту, турецької та інших мов [1]. Для деяких мов реалізовано описи тільки окремих морфологічних підсистем [4].

PC-KIMMO: дворівневий морфологічний аналізатор

PC-KIMMO – це програма, яка здійснює аналіз та/або синтез словоформ, використовуючи дворівневу модель, в якій словоформа подається як відповідність між її формою на лексичному рівні та на поверхневому рівні [2]. Поверхнева форма – це запис словоформи згідно із орфографічними правилами, тобто формування словоформи з основи та аломорфів. Лексична форма словоформи – це узагальнений запис словоформи послідовністю морфем.

У програмному середовищі PC-KIMMO опис морфології природної мови здійснюється створенням двох файлів:

1. Файла правил, в якому визначається абетка та фонологічні (орфографічні) правила.
2. Файла лексикону, в якому містяться лексичні одиниці (слова і морфеми), правила їхньої сполучуваності та відповідні їм морфологічні характеристики.

Наприклад, для англійської мови поверхнева форма словоформи "spies" повинна відповідати її поверхневій формі "`spy+s" так (` – символ наголосу, + символ морфемного шва, 0 – символ нульового елемента):

Лексична форма:	`	s	p	y	+	0	s
Поверхнева форма:	0	s	p	i	0	e	s

Правила повинні описувати відповідність між символами двох рівнів `:0, y:i, +:0, та 0:e. Наприклад, дворівневе правило для встановлення відповідності між y:i буде таким:

y:i => @:C___+:0 (1)

Генератор та аналізатор є основними функціональними частинами PC-KIMMO. Генератор на вході отримує лексичну форму, застосовує до неї фонологічні (орфографічні) правила і повертає відповідну поверхневу форму. Аналізатор на вході одержує поверхневу форму, застосовує до неї фонологічні (орфографічні) правила, переглядає лексикон і повертає відповідну лексичну форму з набором відповідних морфологічних [3].

Правила і лексикон опрацьовують, використовуючи скінченні автомати. Наприклад, дворівневе правило (1) для його використання в PC-KIMMO повинно бути перетворене на таку таблицю переходів скінченного:

	@ C	y i	+ 0	@ @
1:	2	0	1	1
2:	2	3	2	1
3.	0	0	1	0

Розроблення файлів правил та лексикону для опису морфології української мови.

Для реалізації опису морфології української мови в інструментальному середовищі PC-KIMMO необхідно розробити файл правил та файл лексикону.

Потрібно передбачити, що символи кирилиці не підтримуються в PC-KIMMO і буде використовуватись така відповідність між символами:

а-a, б-b, в-v, г-g, ґ – G, д-d, е-e, є-E, ж-Z, з-z, и-y, і-i, ї-I, й-j, к-k, л-l, м-m, н-n, о-o, п-p, р-r, с-s, т-t, у-u, ф-f, х-h, ц-c, ч-C, ш-S, щ-H, ю-U, я-A, ь-B.

Загальна структура файла правил складається із списку оголошень ключових слів. У файлі правил використовується такий перелік ключових слів: ALPHABET, NULL, ANY, BOUNDARY, SUBSET, RULE і END.

1. Перший рядок файла правил визначає символи абетки та окремі.
   ALPHABET{символи абетки української мови}
   a b v g G d e E Z z y i I j k l m n o p r s t u f h c C S H U A B ' - +
   Символ '+' – морфемний шов необхідно також долучити до абетки.
2. В наступних рядках визначається символи NULL (нульовий), ANY (символ підстановки автомата), і BOUNDARY (межа словоформи):
   NULL 0
   ANY @
   BOUNDA
   RY #

Для компактнішого написання файла правил використовується секція SUBSET. Назва груп символів починається з великої літери.

```
SUBSET Ss   y u e  {и, у, е}
SUBSET Apost   A U I E {я, ю, ї, є}
SUBSET Ie   e a i {е, а, і}
```

3. Після пропуску одного рядка починається опис правил перетворень. Перше правило (перший скінченний автомат) повинен визначити задану за замовчуванням відповідність символів поверхневого рівня між собою. Будується умовний автомат, який має тільки один стан і один перехід для кожного поверхневого символу.

Оскільки на поверхневому рівні можуть бути використані 33 символи, таблиця переходів скінченного автомата буде мати 33 стовпчики, потрібно також додати символ ANY, @ (але символ + додавати не потрібно – він використовується тільки на лексичному рівні). У результаті отримаємо таку таблицю переходів скінченного автомата:

```
RULE  "Surface  Characters"   1  34
    a b v g G d e E Z z y i I j k l m n o p r s t u f h c C S H U A B @
    a b v g G d e E Z z y i I j k l m n o p r s t u f h c C S H U A B @
 1: 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

Наступні правил дають змогу врахувати особливості морфології української мови.

Для автоматизації побудови скінченних автоматів, які відповідають правилам, використовується програма KGEN. Вхідними даними для цієї програми є правила, записані за певним форматом. У результаті роботи програми отримуємо повні таблиці переходів скінченного автомата для кожного з правил, які необхідні для роботи системи PC-KIMMO. Вхідний файл для програми KGEN під час розроблення морфологічного аналізатора української мови такий:

```
SUBSET Ss  y u e
SUBSET Apost  A U I E
SUBSET Ie  e a i
PAIRS a b v g G d e E Z z y i I j k l m n o p r s t u f h c C S H U A B
      a b v g G d e E Z z y i I j k l m n o p r s t u f h c C S H U A B
PAIRS   ' - - + d
         ' - 0 0 0
RULE s:S => __ Ss
RULE 0:'=> __ Apost
```

```
RULE t:C => __Ss
RULE k:c => __i
RULE e:i => __l #
RULE i:e => __C Ie
RULE 0:C => __U
END
```

Вхідний файл містить перші сім правил, які описують морфологічні явища української мови. Короткий опис правил наведено у табл. 1.

Таблиця 1

Перелік правил морфологічного аналізатора

Правило	Коментар	Приклад
RULE s:S => __ Ss	Чергування ‘с’ і ‘ш’ в основах перед ‘и’, ‘у’, ‘е’	‘писати’ - ‘пишу’
RULE 0:'=> __ Apost	Апостроф	‘пити’ - ‘п’ю’, ‘бити’ - ‘б’ю’
RULE t:C => __Ss	Чергування ‘т’ і ‘ч’ в основах перед ‘и’, ‘у’, ‘е’	‘летіти’ - ‘лечу’, ‘топтати’ - ‘топчеш’
RULE k:c => __i	Чергування ‘к’ і ‘ц’ на закінченні основ перед ‘і’	‘рука’- ‘руці’
RULE e:i => __l #	Чергування ‘е’ і ‘і’ в основі у родовому відмінку множини	‘село’- ‘сіл’
RULE i:e => __C Ie	Чергування ‘і’ і ‘е’ в основах перед ‘ча’, ‘чі’, ‘че’	‘піч’ - ‘печі’ - ‘печам’ - ‘печей’
RULE 0:C => __U	Подвоєння ‘ч’ в основах перед ‘ю’	‘піч’ - ‘піччю’

У результаті роботи програми KGEN отримуємо такий файл правил, який містить перетворені правила у таблиці переходів скінченного автомата:

```
ALPHABET
   a b v g d Ge E Z z y i I j k l m n o p r s t u f h c C S H U A B ' - +
NULL 0
ANY @
BOUNDARY #
SUBSET Ss   y u e
SUBSET Apost   A U I E
SUBSET Ie   e a i

RULE "defaults" 1 34
   a b v g G d e E Z z y i I j k l m n o p r s t u f h c C S H U A B @
   a b v g G d e E Z z y i I j k l m n o p r s t u f h c C S H U A B @
 1: 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

RULE "defaults" 1 6
   ' - - + d @
   ' - 0 0 0 @
 1: 1 1 1 1 1 1

RULE " s:S => __ Ss" 2 3
     s  Ss   @
     S  Ss   @
 1:   2   1   1
 2.   0   1   0
```

```
RULE " 0:'=> __ Apost" 2 3
    0  Apost   @
    '  Apost   @
 1:  2     1   1
 2.  0     1   0

RULE " t:C => __u" 2 3
    t  u   @
    C  u   @
 1:  2  1   1
 2.  0  1   0
RULE " t:C => __Ss" 2 3
    t  Ss   @
    C  Ss   @
 1:  2   1   1
 2.  0   1   0

RULE " k:c => __i" 2 3
    k  i   @
    c  i   @
 1:  2  1   1
 2.  0  1   0

RULE " e:i => __l #" 3 4
    e  l  #   @
    i  l  #   @
 1:  2  1  1   1
 2.  0  3  0   0
 3.  0  0  1   0

RULE " i:e => __C Ie" 3 4
    i  C  Ie   @
    e  C  Ie   @
 1:  2  1   1   1
 2.  0  3   0   0
 3.  0  0   1   0

RULE " 0:C => __U" 2 3
    0  U   @
    C  U   @
 1:  2  1   1
 2.  0  1   0
END
```

Файл лексикону в PC-KIMMO складається з двох частин. У першій частині описуються стани скінченного автомата, а в другій частині – дуги переходів між цими станами. Стани скінченного автомата описуються в частині, яка називається Alternations, а його переходи в частині Lexicons [2]. В Alternations описуються всі назви станів скінченного автомата, який дає змогу

побудувати словоформу. Перші складові цієї частини описують кореневі поняття дерева лексикону. Кожен рядок файла лексикону має такий формат:

1. Ключове слово ALTERNATION.
2. Назва класу Alternation (стан скінченного автомата).
3. Одна або більше назва стану скінченного автомата. Наприклад:

ALTERNATION Begin <назви станів скінченного автомата >

Один з станів скінченного автомата повинен здійснювати перехід до спеціального стану End. Наприклад: ALTERNATION Foo End

Перший рядок файла лексикону для морфологічного аналізатора української мови буде таким:

ALTERNATION Begin prefixVV1 prefixVV2 prefixVV3 prefixVV4 verb/var1 verb/var2 verb/var3 verb/var4 noun/var1/tv noun/var2/tv noun/var3 noun/var4 noun/var1/mi noun/var2/mi noun/var1/mj noun/var2/mj

Після опису можливих станів скінченного автомата необхідно описати переходи між цими станами. Опис кожного з переходів містить три основні складові: (1) символ переходу, рядок символів лексичної форми; (2) назву можливого стану скінченного автомата; (3) будь-який рядок символів, який необхідно подати на вихід при переході між станами скінченного автомата (набір морфологічних характеристик). Фрагменти файла лексикону з описом переходів наведено в табл. 2.

Таблиця 2

Фрагменти файла лексикону

Приклад опису префікса ‘на’:	Приклад опису кореня ‘смаж’ дієслова:
\lf na+ \lx prefixVV1 \alt verb/var1 \fea \gl Dokonanyj vyd_	\lf smaZ \lx verb/var4 \alt suffixVV4 \fea \gl smaZ
Приклад опису кореня ‘фермер’ іменника:	Приклад опису суфікса ‘ити’:
\lf fermer \lx noun/var2/tv \alt endingNV2/tv \fea \gl fermer	\lf yty \lx suffixVV3 \alt endingVInfV3 \fea \gl +yty_
Приклад опису флексії ‘меш’:	Приклад опису флексії ‘ла’:
\lf meS \lx endingVInfV1 \alt End \fea \gl +meS_sg.2 pers. Future	\lf la \lx endingVPastV4 \alt End \fea \gl +la_sg.fem.past

Морфологічний аналізатор української мови, розроблений в програмному середовищі PC-KIMMO, дає змогу здійснювати морфологічний аналіз та синтез певного набору іменників та дієслів. У табл. 3 та 4 наведені флексії та суфікси, які дають змогу встановити такі морфологічні характеристики іменників, як група, відміна, відмінок та число [3].

Таблиця 3

Група	Відмінок	Закінчення однини	Закінчення множини	Закінчення однини	Закінчення множини
		Перша відміна		**Друга відміна**	
тверда	Н.	-а	-и	-0/--	-и/а
	Р.	-и	--	-а	-ів/--
	Д.	-і	-ам	-у	-ам
	З.	-у	--	-о	-ів/и/а
	О.	-ою	-ами	-ом	-ами
	М.	-і	-ах	-і/ові	-ах
	Кл.	-о	-и	-е/у/о	-и/а
м'яка	Н.	-я	-і	-е/я/--	-і/ї/я
	Р.	-і	--	-я/ю	-ів/їв/ь
	Д.	-і	-ям	-ю	-ям
	З.	-ю	--	-я/е/--	-ів/ї/я
	О.	-ею	-ями	-ем/єм/ям	-ями
	М.	-і	-ях	-і/ї/ю	-ях
	Кл.	-е	-і	-е/ю/я	-і/ї/я
мішана	Н.	-а	-і	--	-і/а
	Р.	-і	-ей	-а	-ів/--
	Д.	-і	-ам	-у	-ам
	З.	-у	-ей	-а/е/--	-і/ів/а
	О.	-ею	-ами	-ем	-ами
	М.	-і	-ах	-і/у	-ах
	Кл.	-е	-і	-е	-і/а

Таблиця 4

Відмінок	Закінчення однини	Закінчення множини	Закінчення однини	Закінчення множини
	Третя відміна		**Четверта відміна**	
Н.	--	-і	-а/я	-а
Р.	-і	-ів/ей	-и	--
Д.	-і	-ям/ам	-і	-ам
З.	-ю	-ів/і	-я	-а
О.	-ом	-ами/ями	-ам/ям	-ами
М.	-і	-ах/ях	-і	-ах
Кл.	-е/и	-і	-я/а	-а
Суфікси	-	-	-ат, - ен, -ят	-ат, - ен, -ят

Загалом дієслова, аналіз яких забезпечує цей морфологічний аналізатор, складаються з основи і закінчення. У табл. 5 наведено перелік флексій, суфіксів та префіксів, які визначать час, особу та число дієслів [3].

Таблиця 5

Час	Особа Число	Дієслова із закінченням 'ять'	Дієслова із закінченням 'уть'	Дієслова із закінченням 'ють'	Дієслова із закінченням 'ать'
1	2	3	4	5	6
Теперішній	1р, sg	-у	-у	-ю	-у
	2р, sg	-иш	-иш	-єш	-иш
	3р, sg	-ить	-ить	-є	-ить

Продовження табл. 5

1	2	3	4	5	6
	1p, pl	-имо	-имо	-ємо	-имо
	2p, pl	-ите	-ите	-єте	-ите
	3p, pl	-ять	-уть	-ють	-ать
Минулий	1p, sg	-в	-в	-в	-в
	2p, sg	-в	-в	-в	-в
	3p, sg	-в(M),-ла(F), -ло(N)	-в(M),-ла(F), -ло(N)	-в(M),-ла(F), -ло(N)	-в(M),-ла(F), -ло(N)
	1p, pl	-ли	-ли	-ли	-ли
	2p, pl	-ли	-ли	-ли	-ли
	3p, pl	-ли	-ли	-ли	-ли
Майбутній	1p, sg	-му	-му	-му	-му
	2p, sg	-меш	-меш	-меш	-меш
	3p, sg	-ме	-ме	-ме	-ме
	1p, pl	-мемо	-мемо	-мемо	-мемо
	2p, pl	-мете	-мете	-мете	-мете
	3p, pl	-муть	-муть	-муть	-муть
Інфінітив		-ити	-ати	-ити	-ити
Суфікси		-ити, и	-ати, и	-ити, и	-ити, и
Префікси		На -	На -	Ви-	За-

Таблиця 6

Приклади використання дворівневого морфологічного аналізатора української мови

Словоформа	Результати морфологічного аналізу
‘пишу’ (зміна приголосної ‘с’->’ш’ перед голосною ‘у’)	PC-KIMMO>recognize pySu pysu [pys_+u_1 sg present] PC-KIMMO>
‘піччю’ (подвоєння приголосної ‘чч’->’ч’)	PC-KIMMO>recognize piCCU piCU [piC+U_Possesive case,sg]
‘смажитимемо’	PC-KIMMO>recognize smaZytymemo smaZytymemo [smaZ+yty_+memo_pl.1 pers. future] PC-KIMMO>
‘вип’ю’ (префікс ‘ви’, корінь ‘п’, флексія ‘ю’)	PC-KIMMO>recognize vyp’U vy+pU [Dokonanyj vyd_p_+U_1 sg present]
‘випити’	PC-KIMMO>recognize vypyty vy+pyty [Dokonanyj vyd_p+yty__infinityv]
‘телятах’	PC-KIMMO>recognize telAtah telAtah [tel+At_+ah_locative case.pl]
‘лошати’	PC-KIMMO>recognize loSaty loSaty [loS+at_+y_Genitive case,sg]
‘печі’ (чергування голосної ‘і’->’е’)	PC-KIMMO>recognize peCi piCi [piC+i]
‘сіл’ (чергування голосної ‘е’->’і’)	PC-KIMMO>recognize sil sel [sel_sg]

Висновки

Розроблення морфологічного аналізатора української мови в програмному середовищі PC-KIMMO є можливим.

Розроблення морфологічного аналізатора полягає у створенні файла правил і файла лексикону.

Доцільно використовувати розроблений морфологічний аналізатор у навчальному процесі.

Доцільно продовжити розроблення морфологічного аналізатора української мови у програмному середовищі PC-KIMMO.

1. Гельбух А. Ф. Эффективно реализуемая модель морфологии флективного естественного языка / Диссертация на соискание ученой степени кандидата технических наук. – М., 1994. 2. Antworth E.L. PC-KIMMO: a two-level processor for morphological analysis. Occasional Publications in Academic Computing No. 16. Dallas: Summer Institute of Linguistics, 1990, 273 p. 3. Плющ М.Я., Грипас Н.Я. Граматика української мови в таблицях. – К.: Вища школа. – 2004. – С. 5–55 с. 4. Bear J. A morphological recognizer with syntactic and phonological rules // Proceedings of Coling'86, Association for Computational Linguistics, 1986, p. 272–276. 5. Jurafsky D., Martin J. Speech and Language Processing. – Prentice-Hall, 2000.