УДК 658.012.56

Т. П. Любченко

МОРФОЛОГІЧНА МОДЕЛЬ СЛОВОЗМІНИ ФЛЕКТИВНОЇ МОВИ ТА ЕЛЕКТРОННИЙ ГРАМАТИЧНИЙ СЛОВНИК

1. Вступ

У системах автоматичної обробки тексту (АОТ), зокрема в системах машинного перекладу (МП), типовими є дві операції над словами: лематизація, тобто редукція текстового слова до його словникової форми, та синтез необхідної текстової словоформи від її початкової (словникової) форми. Вирішення зазначених завдань для мов флективного типу, до яких належить і російська мова, спрямовує до необхідності комп'ютерно-лінгвістичного дослідження системи словозміни, побудови парадигматичної класифікації лексики, створення бази даних, що підтримує дану систему словозміни, розробки алгоритмів формування як повної словозмінної парадигми слова, так і конкретної його текстової форми. Необхідність реалізації такої програми на репрезентативному мовному масиві (хоча б у обсязі 150 тисяч словникових одиниць) робить розв'язання цих завдань виключно складним у «ручному» режимі й вимагає розробки спеціальної комп'ютерної технології.

Для російської мови цю проблему в ряді відомих систем МП і АОТ успішно вирішено, однак скористатися цими результатами майже неможливо, оскільки відомі нам програмні продукти, поширені на ринку, є закритими. Викладене й спонукало нас до самостійної розробки технології автоматизованої побудови словозмінної парадигми російської мови, створення електронного граматичного (парадигматичного) словника, який став би інструментом морфологічного аналізу російських текстів у системі МП, що розробляється в Українському мовно-інформаційному фонді Національної академії наук України (УМІФ НАНУ).

Для побудови граматичного словника визначальним фактором є наявність формальної моделі словозміни, що означає встановлення та формалізацію лінгвістичних критеріїв, згідно з якими вся множина слів мови розбивається на певні підмножини, взаємний перетин яких є порожнім і всередині якого словозміна відбувається за однаковими правилами. Підмножини слів з такими властивостями називаються *словозмінними парадигматичними класами*.

Моделювання розподілу множини слів мови на словозмінні парадигматичні класи відбувається у декілька етапів. На першому визначається поняття *парадигматичного типу*, в означенні якого принципову роль відіграють поняття граматичної категорії, граматичного значення та граматичної форми [1].

2. Словозмінні парадигматичні типи та словозмінні параметри російської мови

Уведемо позначення. Нехай L — фіксована мова (флективна)[1], W — множина слів мови L;

P_j ($j = 1, 2, ..., p$) — граматичні класи[2], p — кількість граматичних класів;

$W(P_j)$ — множина слів мови L, яка належить до граматичного класу P_j;

T_i ($i = 1, 2, ..., N$) — парадигматичні (морфологічні) типи, N — кількість парадигматичних (морфологічних) типів;

$W(T_i)$ — множина слів мови L, яка належить до типу T_i;

$\Omega(T_i)$ — множина граматичних значень, що відповідають типу T_i.

За ознакою приналежності до певної частини мови та за додатковими ознаками, які є класифікуючими (не словозмінними) в межах певної частини мови, множину слів W розподіляємо на підмножини, котрі називатимемо *граматичними класами*, таким чином.

Іменники за значенням граматичної категорії «*рід*» (яка в межах цієї частини мови є класифікуючою ознакою) розподіляються на три граматичні класи: іменники чоловічого роду, іменники жіночого та іменники середнього роду; множинні іменники утворюють окремий граматичний клас [2]. Таким чином, іменники складають чотири граматичні класи (позначатимемо їх відповідно P_1, P_2, P_3, P_4).

Дієслова за значенням граматичної категорії «*вид*» (яка не є словозмінною ознакою і розглядається нами як класифікуюча) розподіляються на такі три граматичні класи: дієслова доконаного виду, дієслова недоконаного виду та двовидові дієслова [2].

Усі інші слова з множини W, які не є іменниками або дієсловами, віднесені до своїх граматичних класів за ознакою приналежності до конкретної частини мови (тобто у цьому випадку поняття «*граматичний клас*» збігається з поняттям «*частина мови*»).

Таким чином, у російській мові нами виділено такі граматичні класи: іменники чоловічого роду (P_1), іменники жіночого роду (P_2), іменники середнього роду (P_3), множинні іменники (P_4), ад'єктиви (прикметники+порядкові числівники) (P_5), дієслова доконаного виду (P_6), дієслова недоконаного виду (P_7), двовидові дієслова (P_8), дієприкметники (P_9),

[1] Наводиться на прикладі російської мови.

[2] Граматичний клас — аналог частини мови.

займенники (займенники-іменники) (P_{10}), займенники-прикметники (P_{11}), числівники кількісні (P_{12}), прислівники (P_{13}), вигуки (P_{14}), сполучники (P_{15}), частки (P_{16}), прийменники (P_{17}), предикативи (P_{18}), скорочення (P_{19}).

Отже, $W = \bigcup_{j=1}^{19} W(P_j)$. Вважатимемо омонімію зняттою, а омоніми промаркованими. Тоді

$$W(P_{j_1}) \bigcap W(P_{j_2}) = \varnothing \text{ при } j_1 \neq j_2,\ j_1, j_2 = 1, 2, \ldots, 19.$$

За словозмінними категоріями, що визначають словозмінну парадигму конкретних слів (сукупність граматичних значень та відповідних граматичних форм), вводимо такі *парадигматичні типи*.

Парадигматичний тип, що характеризується граматичними формами, які визначаються граматичними значеннями словозмінних категорій *«число»* та *«відмінок»*, називатимемо *субстантивним парадигматичним типом*[3]:

$$W(T_1) \equiv W^S = \{w_1^S, w_2^S, \ldots, w_{12}^S\}. \quad (1)$$

При цьому граматичні форми $w_i^S = w_i^S(n,k)$[4] визначаються множиною граматичних значень $\Omega(T_1) = \{\omega_i^S\}$, елементами якої є значення граматичних категорій *«число»* та *«відмінок»* (тобто граматичні форми визначаються парами граматичних значень (*число, відмінок*)):

$$\omega_i^S = \{n_1, k_i\},\ \omega_{i+6}^S = \{n_2, k_i\},\ i = 1, 2, \ldots, 6, \quad (2)$$

де n_1 — однина, n_2 — множина; k_i — значення відмінків: k_1 — називний, k_2 — родовий, k_3 — давальний, k_4 — знахідний, k_5 — орудний, k_6 — місцевий.

До субстантивного парадигматичного типу належать усі іменники та займенники-іменники (займенники, які є замінниками іменників — особові займенники *я, ты, он, мы, вы, они* та такі як *кто, что, некто, нечто*).

Зазначимо, що в кожній з граматичних форм конкретна лексема може мати одну або декілька словоформ чи не мати жодної словоформи. За відсутності реалізації лексеми в певній граматичній формі словозмінну парадигму вважатимемо дефектною. Прикладом подібних випадків є відсутність форм множини в іменників singularia tantum, або відсутність форм однини в іменників pluralia tantum. Для урахування факту відсутності словоформ для деяких граматичних значень у формулу опису парадигматичного типу (1) введемо параметр *def*, який називатимемо *параметром дефектності*:

$$W(T_1) \equiv W^S = \{w_1^S, w_2^S, \ldots, w_{12}^S, def\},$$
$$\Omega(T_1) = \{\omega_1^S, \omega_2^S, \ldots, \omega_{12}^S\}, \quad (3)$$

що вказує на номери (числа) граматичних значень, для яких відповідні словоформи відсутні в повній парадигмі; якщо дефектності немає, то за визначенням покладаємо $def = 0$.

Парадигматичний тип, що характеризується граматичними формами, які визначаються множиною граматичних значень $\Omega(T_2) = \{\omega_i^A\}$, елементами якої є значення граматичних (словозмінних) категорій *«рід»*, *«число»* та *«відмінок»*, будемо називати *ад'єктивним парадигматичним типом*[5]:

$$W(T_2) \equiv W^A = \{w_1^A, w_2^A, \ldots, w_{28}^A, def\},$$
$$\Omega(T_2) = \{\omega_1^A, \omega_2^A, \ldots, \omega_{28}^A\}, \quad (4)$$

де граматичні форми w_i визначаються для граматичних значень ω_i^A, поданих трійками (*рід, число, відмінок*). Для форм однини:

$$\omega_i^A = \{g_1, n_1, k_i\},\ \omega_{i+6}^A = \{g_2, n_1, k_i\},$$
$$\omega_{i+12}^A = \{g_3, n_1, k_i\},\ i = 1, 2, \ldots, 6, \quad (5)$$

а для форм множини значення роду нерелевантне:

$$\omega_{i+18}^A = \{n_2, k_i\},\ i = 1, 2, \ldots, 6. \quad (6)$$

У формулах (4)–(6) g_1 — чоловічий рід, g_2 — жіночий рід, g_3 — середній рід, n_1 — однина, n_2 — множина; k_i — значення відмінків: k_1 — називний, k_2 — родовий, k_3 — давальний, k_4 — знахідний, k_5 — орудний, k_6 — місцевий.

Короткі граматичні форми $w_{25}^A, w_{26}^A, w_{27}^A, w_{28}^A$ існують тільки для називного відмінку і визначаються категоріями роду та числа:

$$\omega_{25}^A = \{g_1, n_1, k_1\},\ \omega_{26}^A = \{g_2, n_1, k_1\},$$
$$\omega_{27}^A = \{g_3, n_1, k_1\},\ \omega_{28}^A = \{n_2, k_1\}. \quad (7)$$

До ад'єктивного парадигматичного типу належать прикметники, займенники-прикметники, порядкові числівники, дієприкметники, кількісний числівник *«один»*.

Для врахування можливої дефектності парадигми у деяких прикметників до формули (4) введено параметр дефектності *def*, який має той самий зміст, що і в (3).

Словозміна дієслів характеризується граматичними формами, що визначаються граматичними значеннями категорій *«стан»*, *«час»*, *«число»*, *«особа»*,

[3] Для російської мови субстантивний парадигматичний тип характеризується 12 граматичними формами: 6 відмінків у однині + 6 відмінків у множині. Для української мови субстантивний парадигматичний тип характеризується 14 граматичними формами внаслідок того, що відмінків в українській граматиці 7, а не 6, як в російській.

[4] Граматична форма залежить від значень категорій *«число»* та *«відмінок»*.

[5] Ад'єктивний парадигматичний тип російської мови характеризується 28 граматичними значеннями: 3 значення роду × 6 відмінків однини + 6 відмінків множини + 4 коротких форми (ч. р., ж. р., с. р. та множ.). В українській мові ад'єктивний парадигматичний тип характеризується 24 граматичними значеннями, які визначають повні словозмінні форми (коротких — немає).

«спосіб», *«рід»* (категорія роду релевантна тільки для минулого часу).

Зазначені категорії можуть набувати таких значень:

стан = {активний, пасивний} ($z = \{z_1, z_2\}$);

час = {теперішній, минулий, майбутній} ($t = \{t_1, t_2, t_3\}$);

число = {однина, множина} ($n = \{n_1, n_2\}$);

особа = {перша, друга, третя} ($l = \{l_1, l_2, l_3\}$);

спосіб = {дійсний, умовний, наказовий} ($h = \{h_1, h_2, h_3\}$);

рід = {чоловічий, жіночий, середній} ($g = \{g_1, g_2, g_3\}$).

Отже, *дієслівний парадигматичний тип* описується формулою:

$$W(T_3) \equiv W^V = \{w_0^V, w_1^V, w_2^V, ..., w_{45}^V, def\},$$
$$\Omega(T_3) = \{\omega_0^V, \omega_1^V, \omega_2^V, ..., \omega_{45}^V\}, \quad (8)$$

де w_0 — інфінітив дієслова (збігається з реєстровим словом); $w_1^V, w_2^V, ..., w_6^V$ — представляють граматичні форми активного стану дійсного способу теперішнього часу; граматичні форми w_i^V ($i = 1, 2, ..., 6$) визначаються п'ятіркою категорій (*стан, спосіб, час, число, особа*)[6] при фіксованих значеннях стану ($z = z_1$ — активний стан), способу ($h = h_1$ — дійсний спосіб) і часу ($t = t_1$ — теперішній час):

$$\omega_i^V = \{z_1, h_1, t_1, n_1, l_i\},\ \omega_{i+3}^V = \{z_1, h_1, t_1, n_2, l_i\},\quad i = 1, 2, 3; \quad (9)$$

$w_7^V, w_8^V, ..., w_{10}^V$ — граматичні форми активного стану дійсного способу минулого часу, які визначаються п'ятіркою категорій (*стан, спосіб, час, число, рід*)[7] при фіксованих значеннях стану ($z = z_1$), способу ($h = h_1$) і часу ($t = t_2$ — минулий час):

$$\omega_{i+6}^V = \{z_1, h_1, t_2, n_1, g_i\},\ i = 1, 2, 3; \quad (10)$$

$$\omega_{10}^V = \{z_1, h_1, t_2, n_2\}; \quad (11)$$

$w_{11}^V, w_{12}^V, ..., w_{16}^V$ — граматичні форми активного стану дійсного способу майбутнього часу, що визначаються п'ятірками категорій (*стан, спосіб, час, число, особа*) при фіксованих значеннях стану ($z = z_1$), способу ($h = h_1$) і часу ($t = t_3$ — майбутній час):

$$\omega_{i+10}^V = \{z_1, h_1, t_3, n_1, l_i\},$$
$$\omega_{i+12}^V = \{z_1, h_1, t_3, n_2, l_i\},\ i = 1, 2, 3; \quad (12)$$

w_{17}^V, w_{18}^V — граматичні форми наказового способу, які визначаються четвірками категорій (*стан, спосіб, число, особа*) при фіксованих значеннях стану ($z = z_1$), способу ($h = h_3$ — наказовий спосіб) та особи ($l = l_2$ — друга особа):

$$\omega_{i+16}^V = \{z_1, h_3, t_2, n_i, l_2\},\ i = 1, 2; \quad (13)$$

w_{19}^V, w_{20}^V — дієприслівникові граматичні форми, які визначаються значеннями пар категорій (*стан, час*) при фіксованих значеннях стану ($z = z_1$) (дієприслівники активного стану теперішнього та минулого часу):

$$\omega_{i+18}^V = \{z_1, t_i\},\ i = 1, 2; \quad (14)$$

$w_{21}^V, w_{22}^V, w_{23}^V, w_{24}^V$ — дієприкметникові граматичні форми дієслова, які визначаються категоріями (*стан, час*) (дієприкметники активного та пасивного стану теперішнього та минулого часу):

$$\omega_{i+20}^V = \{z_1, t_i\},\ \omega_{i+22}^V = \{z_2, t_i\},\ i = 1, 2; \quad (15)$$

w_{25}^V — інфінітив пасивної форми дієслова;

$w_{26}^V, w_{27}^V, ..., w_{31}^V$ — граматичні форми пасивного стану теперішнього часу; значення w_i^V ($i = 26, 27, ..., 31$) визначаються п'ятірками категорій (*стан, спосіб, час, число, особа*) при фіксованих значеннях стану ($z = z_2$ — пасивний стан), способу ($h = h_1$ — дійсний спосіб) та часу ($t = t_1$ — теперішній час):

$$\omega_{i+25}^V = \{z_2, h_1, t_1, n_1, l_i\},$$
$$\omega_{i+28}^V = \{z_2, h_1, t_1, n_2, l_i\},\ i = 1, 2, 3; \quad (16)$$

$w_{32}^V, w_{33}^V, ..., w_{35}^V$ — граматичні форми, що визначаються п'ятірками категорій (*стан, спосіб, час, число, рід*) при фіксованих значеннях стану ($z = z_2$), способу ($h = h_1$) і часу ($t = t_2$ — минулий час):

$$\omega_{i+31}^V = \{z_2, h_1, t_2, n_1, g_i\},\ i = 1, 2, 3; \quad (17)$$

$$\omega_{i+35}^V = \{z_2, h_1, t_2, n_2\}; \quad (18)$$

$w_{36}^V, w_{37}^V, ..., w_{41}^V$ — граматичні форми пасивного стану майбутнього часу; значення w_i^V ($i = 36, 37, ..., 41$) визначаються п'ятірками категорій (*стан, спосіб, час, число, особа*) при фіксованих значеннях стану ($z = z_2$ — пасивний стан), способу ($h = h_1$ — дійсний спосіб) і часу ($t = t_3$ — майбутній час):

$$\omega_{i+35}^V = \{z_2, h_1, t_3, n_1, l_i\},$$
$$\omega_{i+38}^V = \{z_2, h_1, t_3, n_2, l_i\},\ i = 1, 2, 3; \quad (19)$$

w_{42}^V, w_{43}^V — граматичні форми пасивного стану наказового способу, що визначаються четвірками категорій (*стан, спосіб, число, особа*) при фіксованих значеннях стану ($z = z_2$), способу ($h = h_3$ — наказовий спосіб) та особи ($l = l_2$ — 2-га особа) :

$$\omega_{i+41}^V = \{z_2, h_3, n_i, l_2\},\ i = 1, 2; \quad (20)$$

w_{44}^V, w_{45}^V — відповідно дієприслівникова та дієприкметникова граматичні форми пасивного стану дійсного способу минулого часу (визначаються категоріями *стан, спосіб* та *час*: $z = z_2$, $h = h_1$, $t = t_2$):

$$\omega_{44}^V = \{z_2, h_1, t_2\},\ \omega_{45}^V = \{z_2, h_1, t_2\}; \quad (21)$$

def — параметр дефектності.

[6] Категорія роду не є релевантною для форм теперішнього і майбутнього часу.

[7] Для форм минулого часу категорія особи не є релевантною.

Формули (9)–(21) описують усі основні можливі граматичні форми синтетичних дієслівних форм, які можуть бути властиві парадигмі дієслова (і які зазвичай залучаються розробниками до словозмінної парадигми дієслова). Не включено аналітичні форми, зокрема форми умовного способу, а також зворотні форми дієслова. Зворотне дієслово розглядається нами як самостійне і змінне відповідно до парадигматичного типу, що описується формулами (8)–(21). Для парадигматичного типу дієслів можливі випадки дефектності декількох видів. Усі вони описуються параметром *def*, який вказує номери граматичних форм, для котрих відсутні варіанти словоформ; *def* = 0, якщо дефектність відсутня. Наведемо деякі особливі види дефектності дієслівної парадигми в російській мові: відсутність синтетичних форм майбутнього часу у дієслів недоконаного виду, за винятком дієслова *быть*; відсутність форм теперішнього часу, а також дієприкметників активного та пасивного станів теперішнього часу у дієслів доконаного виду; відсутність багатьох форм безособових дієслів тощо.

Таблиця 1

Парадигматичні типи російської мови

Парадигматичний тип	Граматичні класи	Словозмінні граматичні категорії	Кількість граматичних значень у повній парадигмі
Субстантивний	Іменники, займенники-іменники	*число, відмінок*	12
Ад'єктивний	Прикметники, займенники-прикметники, порядкові числівники, дієприкметники	*рід, число, відмінок*	28
Дієслівний	Дієслова доконаного виду, дієслова недоконаного виду, двовидові дієслова	*стан, час, число, особа, спосіб, рід*	46
Парадигматичний тип кількісних числівників	Кількісні числівники	*відмінок*	6
«Нульовий» парадигматичний тип — незмінювані слова	Прислівники, вигуки, сполучники, частки, прийменники, предикативи	—	1

Парадигматичний тип, що характеризується шістьма граматичними формами, які визначаються категорією відмінка, притаманний кількісним числівникам (крім числівника *один*). Такий парадигматичний тип називатимемо *парадигматичним типом числівників*:

$$W(T_4) \equiv W^C = \{w_1^C, w_2^C, ..., w_6^C\},$$
$$\Omega(T_4) = \{\omega_1^C, \omega_2^C, ..., \omega_6^C\}. \qquad (21)$$

Усі незмінні слова російської мови можуть бути віднесені до одного парадигматичного типу — вони мають єдину форму подання у мові, а саме ту, яку подано в реєстровій частині словника. До незмінних слів належать прислівники, сполучники, прийменники, вигуки, частки, предикативні слова.

Таблиця 1 ілюструє парадигматичні типи російської мови і відношення (відповідність) між парадигматичними типами, граматичними класами та словозмінними категоріями.

3. Словозмінні парадигматичні класи та відношення парадигматизації

Усередині граматичних класів виділяємо парадигматичні класи.

Дамо формальне визначення парадигматичного класу. Довільна лексема x (з урахуванням її словозмінних варіантів) може бути подана у вигляді комбінації незмінної та змінної складових:

$$x = c(x) * f(x), \qquad (23)$$

де $c(x)$ — частина лексеми x, яка в процесі словозміни залишається незмінною (квазіоснова), $f(x)$ — її змінна складова (квазіфлексія), * — конкатенація.

Змінна та незмінна складові можуть мати як нульову довжину, так і являти собою всю лексему. Наприклад, у парадигмах іменників із суплетивними формами множини (*человек, человека, ..., люди, людей, ...*) незмінна частина дорівнює нулю, а змінна частина представлена всіма словоформами. У парадигмах незмінних слів, навпаки, нулю дорівнює змінна частина.

Повна словозмінна парадигма $[x]$ слова x, що належить до граматичного класу (парадигматичного типу T_i), має вигляд:

$$\pi(x) = c(x) * \{f_i(x)\}, \qquad (24)$$

де $f_i(x)$, $i = 0, 1, 2, ..., n(T_i)$ — змінні частини слова (квазіфлексії) у відповідних граматичних формах; причому в деяких із них може існувати більше однієї словоформи. Для означення даного факту введемо параметр кратності граматичної форми $v(w_i(x))$, який задається цілим числом, рівним кількості можливих форм лексеми x у граматичній формі w_i. У загальному випадку:

$$f_i(x) = \bigcup_{l=0}^{v(w_i(x))} f_{il}, \qquad (25)$$

$l = l(i) = 0, 1, 2, ...$ — індекс кількості словоформ у граматичній формі (залежить від номера граматичної форми i); $f_0(x)$ — квазіфлексія початкової форми, яка для іменника конкретного роду відповідає словоформі називного відмінка однини, для

дієслова — його інфінітиву, для прикметника — словоформі чоловічого роду називного відмінка однини тощо; $n(T_i)$ — кількість граматичних форм у парадигматичному типі T_i.

Покладемо

$$F = \bigcup_{x \in W} \left(\{f_0(x)\}, \{f_{1l}(x)\}, ..., \{f_{n(T_i)l}(x)\} \right) \equiv$$
$$\equiv \{f_{jl}^1, f_{jl}^2, ..., f_{jl}^{N_i}\}, \quad (26)$$

$$j = 0, 1, 2, ..., n(T_i),\ l = l(w_j) = 0, 1, 2, ...$$

Тоді

$$F = \bigcup_{k=1}^{N_i} [F]^k, \quad (27)$$

де $[F]^k = \{f^k\} = \{f_{jl}^k,\ j = 0, 1, ..., n(T_i)\}$, $N_i = N(T_i)$, $l = l(w_j)$.

Таким чином, кожна множина $[F]^k$ складається з квазіфлексій слів, які мають у всіх своїх граматичних формах $w_1, w_2, ..., w_{n(Ti)}$ (парадигматичного типу T_i) однакові змінні складові.

Оскільки $[F]^k$ побудовані таким чином, що до них увійшли унікальні набори квазіфлексій, тобто $[F]^i \neq [F]^j$ при $i \neq j$ $(i, j = 1, 2, ..., N_i)$, то для кожного граматичного класу P_i (парадигматичного типу T_i) можна побудувати відношення π_i на декартовому добутку $P_i \times P_i$, яке визначається так:

$$\forall x^1, x^2 \in P_i\ x^1 \pi_i x^2 : x^1 = c(x^1) * f^k,$$
$$x^2 = c(x^2) * f^k,\ f^k \in [F]^k. \quad (28)$$

Це відношення є відношенням еквівалентності, оскільки воно, очевидно, є рефлексивним, симетричним та транзитивним. Назвемо його *відношенням парадигматизації*.

Фактор-множина P_i / π_i є множиною парадигматичних класів граматичного класу P_i (парадигматичного типу T_i). Очевидно, що різні словозмінні парадигматичні класи не перетинаються. Отже P_i є об'єднанням парадигматичних класів: $P = \bigcup_{j=1}^{n} \Pi_j$. До одного парадигматичного класу входять тільки ті слова, які мають однакові набори квазіфлексій для всіх граматичних форм, а відрізняються один від одного лише незмінною складовою $c(x)$. Зрозуміло також, що слова з одного класу еквівалентності, визначеного в такий спосіб, мають і однакові правила словозміни.

Таким чином, для кожного з граматичних класів (парадигматичних типів T_i) будується розбиття на множини слів, що не перетинаються, і які є парадигматичними класами, всередині кожного з яких діють єдині правила словозміни. (Для мов флективного типу це означає однаковість флексій граматичних форм та збіг характеру чергування в основі.)

4. Оператор парадигматизації

Для автоматичної побудови повної парадигми за вихідною (початковою) формою x_0 визначається *оператор парадигматизації*

$$H: x \to [x] = c(x) * \{f_0(x), f_1(x), ..., f_n(x)\} \equiv$$
$$\equiv \{c(x) * f_0(x), c(x) * f_1(x), ..., c(x) * f_n(x)\}, \quad (29)$$

дія якого визначається відношенням $\pi(x^1, x^2)$.

Оператор повної парадигматизації (який діє на множині лексем W) визначається за формулою:

$$H = \sum_{i=1}^{4} H_i \cdot \delta(x; T_i), \quad (30)$$

де

$$\delta(x; T_i) = \begin{cases} 1, & x \in W_{T_i}; \\ 0, & x \notin W_{T_i}. \end{cases} \quad (31)$$

H_i — оператор парадигматизації, який діє на множині лексем відповідного парадигматичного типу $W(T_i)$. На множині лексем кожного з парадигматичних типів діє свій оператор парадигматизації, оскільки кожен із парадигматичних типів характеризується своїм комплексом значень граматичних категорій:

$$H_1: x_0 \to [x]\ \forall x \in W(T^S), \quad (32)$$

$$H_2: x_0 \to [x]\ \forall x \in W(T^A), \quad (33)$$

$$H_1: x_0 \to [x]\ \forall x \in W(T^V), \quad (34)$$

$$H_4: x_0 \to [x]\ \forall x \in W(T^C), \quad (35)$$

$$\forall x \in W_{\Pi_k} \subset W(T^S)\ H_1^k : x_0 \to c(x) * [F]_1^k,$$
$$H_1 = \sum_{k=1}^{Cnt(T^S)} H_1^k \cdot \delta(x, W_{\Pi_k}) \quad (36)$$

де H_1 — оператор парадигматизації, який діє на $W(T^S)$; $Cnt(T^S)$ — кількість парадигматичних класів у множині $W(T^S)$; $[F]_1^k$ — множина наборів квазіфлексій слів, які належать до парадигматичного типу $W(T^S)$ (кількість елементів цієї множини дорівнює кількості парадигматичних класів $Cnt(T^S)$). Функція

$$\delta(x; W_{\Pi_k}) = \begin{cases} 1, & x \in W_{\Pi_k}; \\ 0, & x \notin W_{\Pi_k}. \end{cases} \quad (37)$$

Аналогічно для інших парадигматичних типів:

$$\forall x \in W_{\Pi_k} \subset W(T^A)\ H_2^k : x_0 \to c(x) * [F]_2^k,$$
$$H_2 = \sum_{k=1}^{Cnt(T^A)} H_2^k \cdot \delta(x, W_{\Pi_k}), \quad (38)$$

$$\forall x \in W_{\Pi_k} \subset W(T^V)\ H_3^k : x_0 \to c(x) * [F]_3^k,$$
$$H_3 = \sum_{k=1}^{Cnt(T^V)} H_3^k \cdot \delta(x, W_{\Pi_k}), \quad (39)$$

$$\forall x \in W_{\Pi_k} \subset W(T^C)\ H_4^k : x_0 \to c(x) * [F]_4^k,$$
$$H_4 = \sum_{k=1}^{Cnt(T^C)} H_4^k \cdot \delta(x, W_{\Pi_k}). \quad (40)$$

де H_2, H_3, H_4 — оператори парадигматизації, які діють відповідно на $W(T^A)$, $W(T^V)$ та $W(T^C)$; $Cnt(T^A)$ — кількість парадигматичних класів ад'єктивного парадигматичного типу (в $W(T^A)$), $Cnt(T^V)$ — кількість парадигматичних класів у $W(T^V)$, а $Cnt(T^C)$ — кількість парадигматичних класів у $W(T^C)$; $[F]_2^k$, $[F]_3^k$, $[F]_4^k$ — множини наборів квазіфлексій слів, які належать до відповідного парадигматичного типу (T^A, T^V або T^C).

Таким чином, для кожного із класів $W(T^N)$, $W(T^A)$, $W(T^V)$ та $W(T^C)$ оператор парадигматизації визначається незалежно. На множині $W(T^0)$ немає необхідності визначати цей оператор через те, що для незмінюваних слів $[x] \equiv x_0$.

Оператор H відображає лексему x на її повну парадигму $[x]$, і його реалізовано за допомогою словника квазіфлексій і набору алгоритмів побудови повних словозмінних парадигм для російської лексики. За допомогою парадигматичного словника довільній лексемі приписується її словозмінний тип. Далі з використанням набору алгоритмів побудови повних словозмінних парадигм здійснюється граматична ідентифікація лексеми x. Після цього лексема набуває представлення (23).

Алгоритмічна реалізація оператора H^{-1} здійснює процес лематизації, тобто зведення довільної словоформи до її вихідної канонічної форми.

Викладена вище морфологічна модель складає концептуальну основу для комп'ютерного моделювання та реалізації функції парадигматичних відношень.

5. Висновки

У роботі подано модель словозмінної системи російської мови, запропоновано формальне визначення поняття парадигматичного типу і парадигматичного класу, розроблено словозмінну класифікацію російської лексики, яка (класифікація) придатна та зручна для використання її в електронному словнику, побудовано і програмно реалізовано оператор парадигматизації, котрий однозначно ставить у відповідність кожному російському слову його парадигматичний клас. Це дозволяє одержувати повні словозмінні парадигми для всіх повнозначних змінюваних частин мови російської мови.

Наведені результати апробовано і верифіковано на масиві російської лексики обсягом близько 170 тисяч лексем. Одержано 1590 парадигматичних класів, серед них 526 класів іменників без урахування власних назв, 633 — з урахуванням власних назв, 792 класи дієслів, 97 класів ад'єктивів, 4 класи дієприкметників, 24 класи кількісних числівників та 39 парадигматичних класів займенників.

Створено електронний граматичний словник російської мови [4] (який є аналогом паперового «Грамматического словаря русского языка» А. А. Залізняка [2]).

Передбачено розробку морфологічних баз даних для інших мов, залучених до системи МП, яка розробляється в УМІФ НАНУ. Ці дослідження вже виконуються для англійської, німецької, іспанської мов. Принципи моделювання системи словозміни російської мови знаходять застосування й для мов іншої будови (які відрізняються від мов флективного типу). Звичайно, кожна мова має свої особливості, урахування котрих спонукає до відповідних змін у структурі даних, а також розробки нових алгоритмів і програм. Паралельно зі створенням ЛБД для згаданих мов буде виконуватися розробка алгоритмів та програмних модулів морфологічного (морфо-синтаксичного) аналізу текстів, написаних відповідними мовами.

Список літератури: 1. Лингвистический энциклопедический словарь / Под ред. В. Н. Ярцева.– М., 1990.– 685 с. 2. *Зализняк А. А.* Грамматический словарь русского языка: Словоизменение.– М.: Русский язык, 1978.– 878 с. 3. *Широков В. А.* Елементи лексикографії.– К.: Довіра, 2005.– 304 с. 4. *Грязнухина Т. А., Любченко Т. П., Рабулец А. Г.* Электронная версия грамматического словаря русского языка (А. А. Зализняк) как инструмент автоматического морфологического анализа русского текста // Докл. научн. конф. «Корпусная лингвистика и лингвистические базы данных», Санкт-Петербург, март 2002 г.– С. 63–70.

Поступила в редколлегию 21.05.2006