УДК 81'322.2

**Надія Міщенко, Наталя Щоголева**
Інститут кібернетики імені В.М. Глушкова НАН України, м. Київ

# ПРО ВИЗНАЧЕННЯ ГРАМАТИЧНИХ ХАРАКТЕРИСТИК ЛЕКСЕМ ТЕКСТІВ ДЛЯ ПОШУКУ ТЕРМІНОСПОЛУК



**Подано опис однієї з функцій поліфункціональної лінгвістичної системи LIS, а саме: визначення граматичних характеристик (морфологічних категорій та лем) лексем текстів на основі результатів морфологічного аналізу з використанням кортежів закінчень. Ця функція є складовою процесу пошуку терміносполук у науково-технічних текстах.**

**The function of polyfunctional linguistic system LIS, namely, determining grammatical categories of lexemes based on the results of morphological analysis of texts using corteges of endings is described. The function is a part of the processes of searching for key words composed of lexemes frequently used in scientific and technical texts.**

Функція визначення граматичних характеристик лексем довільних текстів флективними мовами є складовою багатьох процесів їх лінгвістичного обробляння*, зокрема, з метою побудови різного типу словників лексики текстів. У цій праці розглянено як виконують ці функції компоненти розробленої в Інституті кібернетики НАН України поліфункціональної лінгвістичної системи LIS та як використати знайдені граматичні характеристики для виконання іншої функції – пошуку термінів та складання частотного списку терміносполук у незнайомих фахових текстах.

Найзручнішим практичним засобом для дослідження термінології є частотні списки лексики текстів, використовування яких зумовлено високою частотою вживання термінів у фахових текстах (від 20 до 30 відсотків слововживань), вище якої є лише частота вживання службових слів (від 30 до 40 відсотків слововживань). Через те частотні списки, упорядковані за зменшенням частоти вживання лексем у текстах, дають змогу користувачеві відшукувати терміни, що визначають тематику текстів, у гранично вузьких межах, а саме у межах початкових фрагментів частотних списків.

Зауважимо, що терміном може бути не лише окрема лексема, а і словосполука, частота вживання якої у тексті пропорційна частоті вживання окремих її лексем. Частотний список словосполук науково-технічного тексту, серед яких є терміносполуки, будують такою послідовністю дій системи LIS:

а) пошук повнозначних лексем, представлених у незнайомому тексті найчастіше вживаними словоформами;

б) визначення граматичних характеристик найчастіше вживаних лексем та формування специфікацій цих лексем;

в) генерування словника лексем за сформованими специфікаціями;

г) побудова частотного списку словосполук з найчастіше вживаних лексем, представлених у згенерованому словнику.

З самого початку функціювання системи LIS дії а), в) та г) виконували компоненти системи LIS автоматично, а дію б) виконував користувач. Проте спосіб представлення лексем у частотному списку та класифікація лексем за допомогою кортежів закінчень уможливили **автоматизацію** процесу знаходження граматичних характеристик лексем, що є основним предметом розгляду у цій статті.

Комп'ютерне обробляння текстів у системі LIS з метою визначення граматичних характеристик ужитих у текстах лексем, базується на припущенні, що кожну словоформу тексту складається щонайбільше з двох частин: основи та закінчення. У цій роботі будемо називати **закінченням** і кінцеву буквосполуку, яка, крім граматичного закінчення, містить змінні суфікси або змінну частину основи, а решту словоформи будемо називати **основою**. Службові слова мови мають лише основу.

**Кортеж закінчень** – це послідовність закінчень словоформ в однині та множині. Для іменних частин мови закінчення розміщено у порядку переліку відмінків, починаючи з називного однини і закінчуючи місцевим множини, а для дієслів – у порядку переліку осіб та форми інфінітиву. Кортежі, що містять однакові закінчення, розміщені у різних позиціях кортежів, вважаються різними. Відсутність закінчення позначено у кортежах цифрою 0, а нездатність мати закінчення у певній позиції позначено крапкою.

Кожний кортеж має **мнемонічне ім'я** і визначає однойменний **клас лексем**, які набувають закінчень кортежу у відмінюванні чи дієвідмінюванні. В іменах кортежів кодують основні граматичні характеристики класу лексем та деякі додаткові ознаки. Наприклад, ім'я *іжа* кортежу закінчень української мови означає, що ці закінчення належать іменникам жіночого роду з закінченням -а в основній формі. Для виведення імен кортежів користувачеві використовують розширені імена, наприклад, "ім., жін.р., зак. -а".

---

* опрацювання – (і надалі в статті) – *ред.*

Кортеж з іменем *іжа* складають закінчення: {-а, -и, -і, -у, -и, -0, -ам, -и, -ами, -ах}. Цей кортеж визначає клас іменників жіночого роду, яких відмінюють як іменник *множина*.

Кортеж з іменем *іжість* (іменники жіночого роду з закінченням -ість в основній формі) складають закінчення: {-ість, -ості, -ості, -ість, -ості, -остей, -стям, -ості, -остями, -остях}, яких набувають іменники, відмінювані як іменник *послідовність*.

Кортеж *ісе* (іменники середнього роду з закінченням -е в основній формі) складають закінчення: {-е, -а, -у, -е, -а, -0, -ам, -а, -ами, -ах}. Закінчень кортежу *ісе* набувають іменники, відмінювані як іменник *середовище*.

**Позиції закінчень у кортежах** також мають мнемонічні імена, за допомогою яких позначають відповідні відмінки для іменних частин мови та особи – для дієслів. Приклад імен позицій у кортежі для іменних частин мови: *он*, *ор*, *од*, *оз*, *оо*, *ом*, *мн*, *мр*, *мд*, *мз*, *мо*, *мм*. Перша літера імені означає вказівку на число: *о* – однина, *м* – множина, друга літера – це початкова літера назви відмінка.

Кількість різних кортежів кожної з мов науково-технічних текстів – російської та української – близько двох сотень.

Система LIS може вести пошук терміносполук у текстах будь-якою мовою, що послуговується кирилицею чи латиною, якщо складено опис (специфікацію) граматики спеціальною **формальною метамовою Dual**, яка входить до складу системи LIS. Цей процес творчий і єдиний, що його виконує користувач, який володіє мовою.

**Специфікація граматики мови** – це текст, який містить алфавіт мови, мнемонічні імена відмінків для іменних частин мови, осіб – для дієслів, імена кортежів закінчень та самі кортежі. Трансформація тексту граматичної специфікації у структури даних комп'ютера виконується компонентом MORF, після чого компонент FEST системи LIS готовий до виконання морфологічного аналізу текстів відповідною мовою без словників.

Процес налаштування системи LIS на певну мову містить також побудову компонентом CON словника службових слів за специфікацією цих слів, сформованою користувачем за допомогою системи LIS. Зокрема, специфікації прийменників, що містять відмінки керованих іменників, зазвичай формує користувач, що легше, ніж створювати спеціальну програму для визначення таких відмінків.

Структури даних, побудовані за специфікацією граматики, у подальшому будемо називати також **граматикою.** Отже, граматика – це кілька масивів і таблиць, серед них масив закінчень, представлення закінчень у вигляді дерев, у вершинах яких знаходяться останні літери закінчень, числове представлення кортежів, списки омонімів закінчень тощо.

Висока частота входжень повнозначних словоформ у текст є необхідною умовою для автоматичного розпізнавання граматичних характеристик таких лексем. Чим менша частота входжень словоформ, тим суттєвіша участь користувача у цьому процесі. Інформацію про частоту входження словоформ у текст надають **частотні списки основ словоформ**, яких формують за результатами морфологічного аналізу текстів, якого виконав компонент FEST, з використанням граматики і словника службових слів.

Згідно з результатами експериментів більшість службових слів мають найвищу серед усіх словоформ частоту вживання. Використання словника службових слів у процесі морфологічного аналізу забезпечує розміщення у частотному списку лише повнозначної лексики, оскільки під час морфологічного аналізу службові слова, знайдені у словнику, не розглядають, а основу кожної словоформи, не знайденої у словнику, заносять у так званий **первинний список основ** з посиланнями на місце входження словоформи у текст та на адресу закінчення у масиві закінчень, якщо закінчення розпізнали. Зауважимо, що використання словника службових слів значно скорочує первинний список основ.

За первинним списком основ, одержаним після оброблення всього тексту, компонент PHRASE формує **частотний список основ**, розміщених у порядку зменшення частоти їх вживання. Кожну основу супроводжує список закінчень словоформ з цією основою, кількість і відсоток входжень словоформ у текст та список адрес входжень. Часто вживану словоформу, яка у всіх входженнях має однакове закінчення, подають у частотному списку цілком і трактують у подальшому оброблянні як незмінюване слово.

Початок частотного списку основ з закінченнями містить представлення найчастіше вживаних лексем. Вони є вхідними даними компонента GENS, який автоматично визначає граматичні характеристики лексем, якщо закінчення при основах покривають усі позиції відповідних кортежів. Для визначення граматичних характеристик рідше вживаних лексем потрібна допомога користувача або збільшення обсягу аналізованих текстів.

Кількість найчастіше вживаних лексем залежить від характеру та обсягу аналізованих текстів. Так, наприклад, у вузькофахових статтях кількість найчастіше вживаних лексем обмежують двома-трьома десятками, але вони можуть покривати до 30 відсотків слововживань. Серед найчастіше вживаних лексем трапляються також загальновживані, які можуть входити у терміносполуки.

**Визначити граматичні характеристики** у запропонованому контексті означає: для кожної лексеми, представленої у частотному списку основою зі списком закінчень її словоформ, знайти кортеж закінчень, які збігаються з тими, що при основі. Якщо такий кортеж знайдеться, то його мнемонічне ім'я і подає шукані граматичні характеристики.

**Лему** як основну форму лексеми визначають за закінченням у відповідній позиції знайденого кортежу закінчень.

Для пошуку імені кортежу компонент GENS використовує, крім фрагменту частотного списку з найчастіше вживаними основами, списки омонімів закінчень, які є частиною граматики. **Список омонімів**

**певного закінчення** – це список імен кортежів, у яких трапляється це закінчення, а при кожному імені кортежу – список імен позицій кортежу, де міститься закінчення. Якщо закінчення трапляється у кортежах лише один раз, то сприймаємо його як один омонім.

Як приклад, наведімо фрагменти списків омонімів закінчень, що трапляються у поданих вище кортежах з іменами *іжа* та *ісе*:

| | | | |
|---|---|---|---|
| **-о**: | *іжа* мр; *ісе* мр; | **-а**: | *іжа* он; *ісе* ор, мн, мз; |
| **-ам**: | *іжа* мд; *ісе* мд; | **-ами**: | *іжа* мо; *ісе* мо; |
| **-ах**: | *іжа* мм; *ісе* мм; | **-е**: | *ісе* он, оз; |
| **-и**: | *іжа* ор, мн, мз; | **-і**: | *іжа* од; **-у**: *іжа* оз. |

Наприклад, фрагмент списку омонімів закінчення -а містить чотири омоніми: у кортежі з іменем *іжа* у називному відмінку однини і в кортежі *ісе* у родовому відмінку однини та називному і знахідному відмінках множини.

Очевидно, що до списку омонімів кожного закінчення ім'я кортежу, у якому це закінчення трапляється, входить лише один раз. З іншого боку, списки омонімів кількох закінчень містять те саме ім'я кортежу, якщо ці закінчення складають цей кортеж. Отже, ім'я кортежу закінчень лексеми, заданої у частотному списку основою та множиною закінчень, слід шукати серед спільних імен кортежів у списках омонімів усіх закінчень множини.

Найкращий випадок, коли спільне ім'я кортежу єдине, тоді визначення граматичних характеристик не потребує участі людини. Припустімо, що лексему *множина* подано у верхній частині частотного списку у такому вигляді:

множин (-а, -ах, -о, -ами, -і, -у, -ам, -и).

Тоді з наведених вище фрагментів списків омонімів буде вибрано єдине спільне для всіх закінчень ім'я кортежу *іжа*. Цей ідеальний випадок ілюструє ідею пошуку граматичних характеристик.

На практиці списки омонімів багатьох закінчень містять десятки омонімів, а список закінчень при основі у частотному списку буває неповним. Тоді для однієї основи можна знайти більше, ніж один кортеж, з кількох причин. По-перше, коли основа спільна для різних лексем, що мають різні кортежі закінчень. По-друге, коли різні кортежі містять однакові закінчення, які займають у кортежах різні позиції. По-третє, якщо словоформа часто трапляється у тексті лише в одному чи у двох відмінках, то закінчень може бути замало для однозначної ідентифікації кортежу. У деяких випадках користувач може уточнити інформацію у частотному списку.

**Генерування словоформ за основою та вибраними кортежами** здійснює компонент GENW з метою візуального контролю користувачем за іменами вибраних кортежів. Проаналізувавши результат генерування, користувач викреслює імена зайвих кортежів. А залишені імена подають граматичні характеристики лексем та однозначно визначають їх леми.

Результат роботи компонента GENS подають у вигляді текстового файлу, що містить послідовність специфікацій найчастіше вживаних лексем метамовою Dual у міру розпізнавання їхніх граматичних характеристик.

Приклади специфікацій лексем з наведеними вище кортежами закінчень:

```
множин      => * : іжа;
послідовн   => * : іжість;
середовищ   => * : ісе;
```

Специфікація лексеми *послідовність* може містити імена двох кортежів, другий для прикметника *послідовний* (якщо він часто трапляється), оскільки основи їх збігаються.

Файл, що містить специфікації найчастіше вживаних лексем, є основою для генерування комп'ютерного словника найчастіше вживаних лексем компонентом CON, який може згенерувати окремий словник лексем або розширити цими лексемами словник службових слів, залежно від запланованого користувачем способу використання словників.

**Частотний список словосполук** із найчастіше вживаних лексем формує компонент PHRASE на основі результатів повторного морфологічного аналізу тексту з використанням граматики та розширеного словника лексем. У цьому випадку компонент FEST – виконавець аналізу, формує первинний список відшуканих у словнику словоформ із вказівками на узгодження іменників із сусідніми словоформами. Словосполуки у частотному списку розташовано у порядку зменшення їх довжини, а в межах словосполук однієї довжини – у порядку зменшення частоти їх вживання.

Результати численних експериментів з текстами різної тематики дають змогу стверджувати: переважна більшість часто вживаних словосполук є **терміносполуками**, право остаточного вибору яких належить користувачеві системи.

Реалізований у системі LIS спосіб визначення граматичних характеристик лексики текстів уможливлює пошук терміносполук у фахових текстах незнайомої тематики різними мовами. Настроюють систему на конкретну мову автоматично за специфікаціями граматики мови та службових слів формальною метамовою, яка входить до складу системи.

Запропонований спосіб можна бути застосувати для автоматизованої побудови словників лексики будь-якого тексту великого обсягу, який забезпечує високу частоту входжень майже всіх лексем тексту.

На шляху до автоматичного пошуку граматичних характеристик лексики текстів стоять на заваді дві основні проблеми: хибна омонімія закінчень, оскільки морфологічний аналіз текстів для пошуку найчастіше вживаних лексем виконується без словника повнозначних слів, та високий ступінь омонімічності закінчень, що унеможливлює однозначний автоматичний вибір імені релевантного кортежу.