УДК 681.3.06

А. Г. КАН, В. П. ЯРМОЛЮК

АВТОМАТИЗАЦІЯ ОРФОГРАФІЧНОГО КОНТРОЛЮ В СИСТЕМАХ ПЕРЕРОБКИ ТЕКСТОВОЇ ІНФОРМАЦІЇ

В останні роки зросла увага до принципів організації лінгвістичних баз даних і, зокрема, до методів створення, зберігання та поповнення різного виду словників [1, 8]. Це пояснюється як зростанням інформаційних потоків, що вимагають оперативної обробки за допомогою ЕОМ, так і розвитком засобів обчислювальної техніки і зв'язку, прогресом в області архітектури, технології та математичного забезпечення обчислювальних систем.

Було багато спроб створення словників для машинного перекладу, компресії текстової інформації тощо. З появою у поліграфії автоматизованих систем переробки текстової інформації (АСПТІ) і їх швидким розвитком назріла необхідність у межах бази даних АСПТІ організувати орфографічний словник для контролю текстів, які вводяться в ЕОМ з метою виявлення помилок клавіатурного процесу. Контроль здійснюється шляхом порівняння конструкцій (слів, словоформ, графем, морфем) тексту зі зразками, заданими орфографічним словником, що зберігається в пам'яті ЕОМ.

Введення орфографічного контролю в АСПТІ дасть змогу суттєво скоротити тривалість коректурних циклів, збільшити продуктивність праці коректорів і редакторів.

Розв'язання такої задачі пов'язане з рядом технічних труднощів. Основна з них полягає в тому, що обсяг достатньо повного орфографічного словника оцінюється значенням, яке перевищує десятки Мбайт, що робить реалізацію такого словника в пам'яті ЕОМ досить проблематичною.

Тому існуючі АСПТІ, бази даних яких містять орфографічні словники значно меншого обсягу [9], мають невисоку контролюючу здатність.

Деякі спеціалісти [3] вважають, що відзначені труднощі будуть подолані з удосконаленням обчислювальної техніки, тому немає необхідності вдаватися до спеціальних заходів щодо розробки методів стиснення словників. Однак слід зауважити, що темпи росту інформаційних потоків порівнянні з темпами розвитку ЕОМ, а задачі контролю вимагають невідкладного розв'язання.

Пропонуються два напрямки подолання невідповідності між обмеженою ємністю пам'яті ЕОМ і обсягом інформації, яка поступає в АСПТІ [9]. Перший шлях передбачає введення змінних носіїв інформації, а другий — стиснення словників, що зберігаються в базі АСПТІ. Другому напрямку слід віддати перевагу, оскільки він не порушує однорідності структури пам'яті ЕОМ і не позбавляє систему в цілому оперативності.

Ставиться завдання стиснення орфографічного словника без втрати його інформативності таким чином, щоб створена під-

множина мала контролюючу здатність, яка б не поступалася такій для повного словника.

Стиснення відбувається шляхом абстрагування від значення слова або словоформи, аж до їх редукції до окремих буквених сполучень. Подібне вже робилося у працях [4, 7], але оптимальне вирішення, тобто розумний компроміс між ступенем стиснення словника і його контролюючою здатністю, не було досягнуте.

Перш ніж досліджувати контролюючу здатність словника буквених сполучень, на масиві 550 помилкових слів (вибраних з художніх текстів обсягом 0,5 млн. слів і підготовлених операторами-друкарями для вводу в АСПТІ) визначена контролююча здатність орфографічного словника [5]. При цьому виявлено 56,4% помилкових слів. Решта не знайдена з двох причин: по-перше, частина помилок носила семантичний характер, тобто помилки зумовлені заміною одних дозволених слів іншими; по-друге, деякі слова були відсутні в словнику (технічні терміни, біблейські імена, іншомовні слова та ін.).

Знайдене значення контролюючої здатності словника прийняте як верхня межа оцінки контролюючої здатності таблиць буквених сполучень.

Принцип створення таблиць буквених сполучень оснований на виявленні допустимих сполучень, розміщених поряд у словах російської мови.

Таблиці двобуквених сполучень описані в літературі [2], але вони не підходять для аналізу художніх текстів. Тому ми створили таблиці дозволених двобуквених сполучень для початкових, серединних і кінцевих частин слів. За допомогою таблиць виявлено 15% помилкових слів, що звичайно, зовсім недостатньо для ефективного контролю.

Аналіз помилкових слів показав, що 54% орфографічних помилок знаходяться серед перших чотирьох букв, 35% помилок — серед останніх чотирьох і тільки 11% — в середині слів.

Тому ми запропонували контролювати слова за допомогою таблиць трибуквених і чотирибуквених сполучень тільки початкових і кінцевих частин слів. Справді, таблиці трибуквених сполучень дали змогу виявити 31% помилкових слів, а таблиці чотирибуквених сполучень — 53%.

Високу контролюючу здатність чотирибуквених сполучень початкових і кінцевих частин слів можна пояснити тим, що, по-перше, оскільки середня довжина слова не перевищує вісім букв, то більше половини слів контролюється чотирибуквеними сполученнями практично повністю; по-друге, як відзначено у праці [7], максимуми інформації зосереджені на початку та в кінці словоформи, а букви, які знаходяться в серединній частину словоформи, несуть найменшу інформацію.

Для орієнтовного розрахунку обсягу таблиць чотирибуквених сполучень проведено статистичний експеримент, який полягає у виборці випадковим чином сторінок з орфографічного словника та підрахунку кількості груп слів, які мають однакові перші

чотири букви. Число груп становить приблизно 18% загальної кількості проаналізованих слів. Аналогічний розрахунок проведено по оберненому словнику. Групи слів, які мають чотири однакові останні букви, становлять 12% загальної кількості проаналізованих.

Отже, найбільший обсяг чотирибуквених таблиць можна оцінити орієнтовно $0{,}3 \times 4 \times N$ байт, де N — кількість слів орфографічного словника.

Оскільки при буквеному кодуванні (одна буква представляється одним байтом інформації) обсяг словника становить приблизно $8 \times N$ (середня довжина слова вісім букв), то затрати пам'яті ЕОМ на зберігання чотирьох буквених таблиць приблизно у сім раз менші, ніж затрати на зберігання повного орфографічного словника.

Таким чином, для автоматизації орфографічного контролю текстової інформації, яка вводиться в АСПТІ, можна використати таблиці чотирибуквених сполучень початкових і кінцевих частин слів, що дає суттєву економію пам'яті ЕОМ, не зменшуючи при цьому контролюючу здатність порівняно з повним орфографічним словником.

Список літератури: 1. *Белоногов Г. Г., Новоселов А. П.* Автоматизация процессов накопления, поиска и обобщения информации. — М.: Наука, 1979. 2. *Белоногов Г. Г., Фролов Г. Д.* Эмпирические данные о распределении букв в русской письменной речи. — Проблемы кибернетики, 1963, № 9. 3. *Крос Р.-К., Гардж Ж.-Г., Леви Ф.* Синтол — универсальная модель системы информационного поиска. — М.: ВИНИТИ, 1968. 4. *Лебедев Д. С., Гармаш В. А.* О возможности увеличения скорости передачи телеграфных сообщений. — Электросвязь, 1958, № 1. 5. Орфографический словарь русского языка. — М.: Русский язык, 1974. 6. *Пиотровский Р. Г.* Текст, машина, человек. — Л.: Наука, 1975. 7. *Пиотровский Р. Г.* Информационные измерения языка. — Л.: Наука, 1968. 8. *Солтон Дж.* Динамические библиотечные системы. — М.: Мир, 1979. 9. La machine à corriger les fautes d'orthographe. — Caractere, 1975, № 6.

The organization of an orphographical control in automation text-processing systems in discussed in this article. Controlling ability of an orphographical dictionary, and its subsets, composed from solved letter-combinations is estimated.

Стаття надійшла в редколегію 8 квітня 1980 року