СЛОВНИКОВА СТАТТЯ ЯК ЧАСТИНА ЕЛЕКТРОННОГО СЛОВНИКА

(конструктивний та функціональний аспекти)

© Василь Войнов, Костянтин Носів, Сергій Тимофєєв, 2000
Харківський національний університет

Пропонується вирішення проблеми, засноване на **методиці побудови трансляторів** мов програмування. Відзначимо, що попутно виявляються й усуваються помилки сканування. **Основою вирішення проблеми є створення опису структури словникової статті (СС) у вигляді формальної граматики (Ф-грамматики).**

Програма розпізнавання структури СС є синтаксичним розпізнавачем, що налаштовується Ф-граматикою. Форма представлення Ф-граматики може бути різною: варіанти БНФ (тобто граматики, що породжують), синтаксичні графи (тобто граматики, що розпізнають). Принципи створення програми розпізнавача також можуть бути різними.

Однак "лобовою атакою" вирішити проблему не вдається через специфічну **проблему відсутності формального опису** структури СС. Якщо розроблювачі мов програмування обов'язково дають його формальний опис, то укладачі словників цього, як правило, не роблять. Найімовірніше, такі описи і не складалися, а самі автори словників опиралися на не цілком формалізовані правила структурування, причому іноді відступали від них.

Підтвердженням тому служать різні нерегулярності структури СС, а також структури, що виявляються лише на підставі змісту.

Взагалі кажучи, Ф-опис може бути створений у такий спосіб.

На основі первісного перегляду й аналізу декількох СС лінгвіст створює першу версію Ф-опису. Потім, переглядаючи весь словник стаття за статтею, він уточнює Ф-опис.

Недолік цього методу очевидний: він придатний лише для невеликих словників і вимагає великих зусиль від лінгвіста. Фактор великого обсягу ("перехід кількості в якість") робить цей метод непридатним для великих словників.

Таким чином, ми виявили **три проблеми створення Ф-опису**:

1. Проблема нерегулярності структури СС.
2. Проблема смислової залежності структур СС.
3. Проблема великого обсягу словника.

Для вирішення проблеми великого обсягу словника пропонується взяти на допомогу комп'ютер і формулювати Ф-опис в інтерактивному ітераційному режимі.

Суть методу:

1. Створюється *перше наближення* Ф-опису на основі аналізу обмеженої кількості СС (це робить лінгвіст).
2. Пишеться *програма верифікації* структури СС за методикою синтаксично керованих трансляторів (це робить програміст).
3. Через дану програму, як *крізь сито*, просівається весь словник. Затримані СС мають структуру, не підтримуваної поточним Ф-описом (це робить лаборант).
4. Лінгвіст аналізує частину затриманих статей (скільки зможе) і *уточнює Ф-опис*.

5. Програміст *коректує програму* відповідно до уточненого Ф-опису.
6. Ітерації повторюються, починаючи з п.2, поки всі статті не пройдуть верифікацію.

Для вирішення проблеми нерегулярності структури СС і проблеми семантично залежної структури пропонуються наступні методи.

1. *Метод підтримки виключень.*

   Усі нерегулярності описуються формально. Ф-грамматика ускладнюється, але зате стає універсальною. Вихідний словник не зачіпається.

2. *Метод допоміжних позначок.*

   У текст вихідної СС вставляються особливі помітки, що дозволяють ідентифікувати її структуру без апеляції до семантики.

3. *Метод уніфікації структури СС.*

   Здійснюється семантично еквівалентне перетворення СС, що спрощує її структуру. Метод придатний для боротьби як з нерегулярностями структури, так і з семантичною залежністю структури СС.

Зауважимо, що в методах 2 і 3 можна зберігати (для довідок) вихідний образ СС.

Створення машинного словника.

Зрозуміло, що збереження словника у вигляді тексту неефективно для пошуку й інших операцій над словником.

Машинний словник треба представити у вигляді деякої БД. Тут можливі два варіанти:

1) *Спеціальна структура збереження.* Використовується в комерційних продуктах для цілей максимальної ефективності. Спосіб трудомісткий у реалізації.

2) *Універсальна структура збереження*, реалізована в якій-небудь СУБД загального застосування, наприклад, MS ACCESS. Цей варіант кращий для дослідницьких цілей, а також як попередній етап створення комерційного продукту. Використовуючи могутню мову маніпулювання даними, легко будувати різноманітні алгоритми роботи зі словником.

Тепер можна намалювати **загальну картину перетворення паперового словника в електронний**. Паперовий словник сканується FineReader'ом і перетвориться в набір документів Word'a в одній директорії. Потім кожен документ піддається нескладному перетворенню типу представлення у вигляді суцільного тексту замість декількох стовпчиків і вилучаються переноси слів. Далі лаборант запускає користувальницьку макрокоманду Word'a, що і є програмою верифікації структури словникових статей. Програма відкриває черговий невідпрацьований документ, виділяє чергову словникову статтю, і символ за символом її перевіряє. Як тільки буде виявлена невідповідність структури словникової статті до її опису, програма зупиняється з відповідною діагностикою. Якщо це помилка сканування, лаборант її відразу виправляє і робота програми продовжується. Якщо ж трапилася помилка неповноти опису, лаборант вимагає від програми пропустити дану словникову статтю. Програма позначає її в першому рядку спеціальним чином і переходить до опрацювання наступної статті. Перехід від документа до наступного здійснюється автоматично.

Незалежно від роботи лаборанта, лінгвіст і програміст аналізують позначені статті й або коректують формальний опис і програму, або коректують саму статтю.

Коли програма видає повідомлення "Верифікація структури словникових статей успішно завершена", запускається інша програма, що здійснює контроль відсилочних

словникових статей. Попередньо створюється індекс для прискорення роботи. Цим процедура верифікації завершується.

Під час розроблення програми передбачається використовувати методику створення синтаксично керованих трансляторів мов програмування, доповнену засобами об'єктно-орієнтованого програмування. Програму передбачається писати мовою VBA.

Після описаного вище первинного опрацювання просканованого словника й одержання його вивіреного електронного образу здійснюється конвертування словника в базу даних ACCESS. Усі інші дії над словником здійснюються в цій структурі засобами VBA.

Декілька пропозицій стосовно формальних ознак частин словникових статей у одинадцятитомному СУМі:

1. Заголовна частина словникової статті: великі літери напівжирним шрифтом (з верхнім індексом, можливо), а також морфологічна інформація, подана курсивом, до крапки (**АБЕРАЦІЯ**, ї, *ж.*).

2. Тлумачення: арабська цифра з крапкою, курсивна позначка галузі уживання скорочена, тлумачення прямим нежирним шрифтом до крапки включно перед курсивом поданою ілюстрацією (1. *астр.* Позірне відхилення світил від їх справжнього положення на небозводі, викликане рухом Землі по орбіті (річна аберація) або її обертанням навколо осі (добова аберація)).

3. Ілюстративний текст, поданий курсивом, з вказівкою на джерело ілюстративного тексту, поданою у дужках прямим нежирним шрифтом. Ілюстрація завершується закривальною дужкою та крапкою: *Швидкість руху Землі мізерна в порівнянні з швидкістю світла..;тому гадане переміщення зірок незначне., Проте його можна виявити за допомогою астрономічних приладів. Це явище називається аберацією світла* (Цікава фізика, 1950, С.26).

Тлумачень може бути декілька і кожне описується 2 та 3 ознаками.

Словникова стаття може мати й інші формальні позначки вказаного характеру.

У складі тлумачення можлива наявність сполучення заголовної частини з іншими словами. Сполучення подається малими напівжирними літерами і має додаткове тлумачення та ілюстративний текст. Шрифти та ознаки додаткових тлумачення та ілюстрацій позначені вище: **За абеткою** – за порядком літер, прийнятим в абетці (алфавіті). *Словник – книга, в якій за абеткою розставлені тисячі, десятки, а то й сотні тисяч слів* (Перв., III, 1959, С.310).

Формальні ознаки такого типу організуються у деревоподібній графічній структурі, на основі котрої будується алгоритм автоматичного перетворення словникової статі друкованого словника у записи комп'ютерної бази даних.

1. Словник української мови / К.: Наукова думка, 1971. В 11 томах. 2. Широков В.А. Інформаційна теорія лексикографічних систем / К.: Вид-во «Довіра», 1998. –331с.