УДК 681.3.06:51

Паскевич А.Ю., студент

ОСОБЛИВОСТІ РЕАЛІЗАЦІЇ МОВИ ВИСОКОГО РІВНЯ ДЛЯ ЗАПИСУ МАТЕМАТИЧНИХ ТЕКСТІВ

Розглянуто нову практичну реалізацію мови TL та її особливості у порівнянні з версією, що використовувалась в системі САД. Також приведено опис англомовної версії TL, яка вперше реалізована програмно.

New practical implementation of TL language and its features in comparison with the version used in SAD are considered. Description of the English TL version that was software implemented for the first time is given also.

Велика кількість задач, що виникають у сучасній математиці та суміжних з нею прикладних дисциплінах, вимагають застосування комп'ютерних технологій. Внаслідок цього було розроблено та реалізовано чималу кількість різноманітних комп'ютерних математичних служб. Сюди відносяться як системи комп'ютерної алгебри, що застосовуються для числових та аналітичних перетворень, так і системи автоматичного доведення теорем, що здатні проводити логіко–математичні міркування. Також дедалі набуває значення накопичення декларативних математичних знань, тобто формалізованих теорій, що описують різноманітні галузі математики.

В останні часи посилюється тенденція до інтеграції усіх цих напрямків. У зв'язку з цим постає питання про єдину формальну мову, яка була б з'єднуючою ланкою між різними математичними службами, придатну для передачі та зберігання математичної інформації, як декларативного, так і алгоритмічного характеру. Розвиток комп'ютерних мереж та технологій розподілених обчислень роблять цю задачу ще актуальнішою. З іншого боку, така мова має бути орієнтована насамперед на користувача, бути якомога більш наближеною до природньої мови математичних текстів.

Майже тридцять років тому Інститутом кібернетики в Києві під керівництвом В.М.Глушкова була започаткована програма робіт під назвою "Алгоритм очевидності", яка мала втілити в собі усі згадані аспекти. Її складовою частиною була розробка формалізованої мови для запису математичних текстів. Така мова була остаточно сформована до 1974 року і отримала назву TL (Theory Language) [1]. Вона має формальну граматику, але її конструкції повторюють природню мову.

В TL є поняття речення, аналоги підмету, присудку, означень та доповнень, що реалізуються за допомогою так званих утворювачів, тобто комбінацій фіксованих лексем та аргументних позицій. Утворювачі формують конструкції, атрибути, твердження мови. TL–текст може бути структурованим, складатись з розділів, включати теореми, доведення, визначення і т.д. Наведемо типовий приклад TL–тексту.

Лемма 1. Пусть S есть дерево и a принадлежит S. Пусть множество _ S(a) состоит из x таких что <a,x> in S. Положим S(a) конечно и существует конечный принадлежащий S путь. Тогда существует бесконечный принадлежащий S путь.

У 1979 році з'явилась перша програмна реалізація системи САД (Система Автоматизації Доведень), яка включала у себе декілька методів пошуку логічного виводу у теоріях першого порядку, обробку рівностей, деякі евристичні методи доведення. У тому числі, в САД було реалізовано певну підмножину мови TL.

Ця стаття присвячена альтернативній реалізації цієї мови, що також носить практичний характер. Ця спроба має за мету відновити вже існуючі досягнення у цьому напрямку, а в окремих випадках й вдосконалити їх, або розглянути під іншим кутом зору. На сьогоднішній день вона залишається російськомовною, але може бути легко перекладена українською мовою. Реалізовано три основних блоки, що не потребують апарату логічного виводу: лексико–синтаксичний контроль, введення у мову нових утворювачів та перевірка декларованості змінних. Вкажемо їх основні особливості.

Лексичний аналіз. Для розв'язання проблеми флексій був застосований префіксно–зразковий підхід. При введенні нового слова користувач вказує його незмінну основу та максимальну можливу кількість літер у закінченні. Таким чином, зразок "*множеств???*" буде співставлений слову "*множество*" у будь–якому числі та відмінку. Крім того, в мову було введено поняття коментаря.

Синтаксичний аналіз. Сюди відносяться як аналіз окремого речення, так і контроль структури тексту, оскільки й те, й інше задається граматикою мови. За основу було узято граматику версії, реалізованої в САД.

Принциповою модифікацією є шаблонне подання утворювачів. Кожен утворювач відповідає деякому шаблону, який задає його тип (конструкція, твердження, символьний утворювач) та схему, тобто розташування лексем та позицій. Звісно, кількість шаблонів, що застосовуються у природній мові, дуже обмежена (у програмі реалізовано дев'ятнадцять). У граматику введені відповідні продукції для кожного шаблону, що дозволяє їй не залежати від окремих утворювачів.

При такому підході, лексичний аналізатор має розпізнавати шаблон утворювача за лексемами. В поточній версії розпізнавання відбувається за першою лексемою — лідером. Таким чином, лексема–лідер має бути унікальною у словнику мови.

Більшість внесених у граматику змін походять з прагнення уникнути синтаксичних невизначеностей, що потребували б вирішення на вищих рівнях обробки. Потрібно відмітити, що окремі невизначеності провокуються шаблонним поданням утворювачів. Найбільш помітна з цих "вимушених" модифікацій така: перед правими атрибутами необхідно вживати прийменник **С**. Це не забороняється правилами природньої мови ("*гомоморфизм с ядром K*"), але іноді виявляється недоречним: порівняйте "*слово длины l*" та "*слово с длиной l*".

Окремі правила граматики з попередньої реалізації, що дозволяли математично некоректні вислови, були замінені жорсткішими. Насамперед, це стосується однозначних конструкцій, тобто таких, що описують не клас сутностей, а одиничний об'єкт. Такі вирази, як "*некоторый порядок группы*", в описаній версії не дозволяються.

Крім того, граматика TL була доповнена новими конструкціями, наприклад зв'язкою **ГДЕ** наприкінці речень.

Щодо структурного контролю, структурні та змістовні роздільникі вже не зберігаються окремо, а подані в граматиці. Такий підхід дозволяє розв'язати наступну проблему. У деяких змістовних розділах жорстко регламентовані типи та порядок речень. Наприклад, формулювання теореми складається з послідовності припущень, що завершуються твердженням. На відміну від САД усі такі вимоги "зашиті" у граматику. Необхідна синонімічність роздільників ("*теорема*", "*лемма*", "*утверждение*") легко забезпечується на лексичному рівні.

Визначення нових понять. При введенні нового терміну у системі САД користувач мав самостійно доповнювати граматику мови відповідною продукцією. Оскільки мова TL має бути самодостатнім засобом запису математичних текстів, такий підхід є суттєво недоречним. Тому в новій версії реалізовано введення нових понять через визначення у TL–тексті.

Цей механізм легко здійснюється завдяки шаблонному поданню утворювачів, який дозволяє взагалі не змінювати граматику мови. У граматиці описано варіанти визначень для кожного шаблону конструкцій, тверджень та символьних утворювачів. Нововведений утворювач одразу включається у словник мови, причому дозволяються рекурсивні визначення, тобто використання нового поняття у правій частині дефініції.

Для подання однозначних конструкцій було введено додатковий тип утворювача. Однозначні конструкції не визначаються безпосередньо, а утворюються зі звичайних за допомогою спеціального твердження "*ОПРЕДЕЛЯЕТСЯ ОДНОЗНАЧНО*". Воно може слідувати одразу після визначення або бути заключенням теореми.

Атрибути також не визначаються, а реалізуються правилами граматики через утворювачі тверджень та однозначних конструкцій.

Декларованість змінних. Правила декларування та області дії змінних реалізовані у відповідності з [1]. Введені додаткові можливості при декларуванні змінних у кон'юктному ланцюжку. Так, наприклад, "*Пусть X есть множество и Y есть подмножество X.*" є зараз допустимим виразом. Планується також вдосконалити визначення області дії змінних, декларованих у змістовних розділах.

Англомовна версія TL. Поряд з реалізацією російськомовного TL була розроблена та втілена граматика англомовної версії. Теоретичні дослідження цього питання вже проводились [3], але вони стосувались більш автоматичного перекладу англійською мовою, ніж її розпізнавання. Розроблена англомовна граматика є першою, що була реалізована програмно.

Завдяки певній подібності англійської та російської мов, а також простоті виражальних засобів TL, зміни, що необхідні для переходу до англомовного варіанту, мають характер радше локальний, ніж глобальний. Структура граматики залишається незмінною, а основних розбіжностей лише три.

По–перше, у англійській мові, при утворенні тверджень, лідерам r–утворювачів, заснованих на прикметниках, передує дієслово **IS** (тобто "*is*", "*are*" або "*be*"): "*number _ N is negative*". Цього не відбувається, якщо утворювач заснований на дієслові: "*number _ N divides M*". Крім того, у лівих атрибутах **IS** взагалі не використовується: "*negative number _ N*". Програма синтаксичного контролю розрізняє "прикметникові" і "дієслівні" r–утворювачі та забезпечує їх коректну обробку.

По–друге, ліві атрибути з однією або більше позиціями розташовуються не попереду лідера конструкції, а після нього: "*number _ N bigger than 5*". На відміну від російської мови, розташування зліва неприпустиме, тому граматика була відповідно модифікована.

По–третє, необхідно встановити правила вживання артиклів. Оскільки в TL артиклі не мають змістовного навантаження, вони, як правило, ігноруються. Але в твердженні типу "*set _ W is a subset of T*" наявність артикля суттєва — він запобігає синтаксичній невизначеності. Без артикля, r–утворювач після **IS** може виявитися як твердженням, так і лівим атрибутом конструкції.

Особливості програмної реалізації. Нагадаємо, що система САД була реалізована на машинах серії ЄС–ЕОМ мовою програмування PL–1. В якійсь мірі це й спричинило спробу альтернативної реалізації замість переносу вже існуючої системи. Розроблена програма написана мовою програмування C. Власне блок синтаксичного контролю створений за допомогою генератора парсерів Bison. За операційну платформу була обрана OS Linux (UNIX). Програма розроблена так, щоб її перенесення на інші платформи або адаптація під можливі нові технології коштували якнайменше сил та часу.

Заключення. Безумовно, розглянута реалізація не є остаточною. Передбачається подальше накопичення мовних засобів, забезпечення певного поліморфізму мови, розробка зовнішньої бази понять.

Робота проводилась у рамках проекту INTAS 96–0760.

Автор висловлює подяку М.К.Мороховець та О.В.Лялецькому, чиї поради та досвід підтримували і направляли цю роботу.

ЛІТЕРАТУРА

[1] О формальном языке для записи математических текстов / В.М.Глушков, К.П.Вершинин, Ю.В.Капитонова и др. // Автоматизация поиска доказательства теорем в математике. — Киев: ИК АН УССР, 1974.

[2] О понятии корректности текста в языке TL / К.П.Вершинин // Математические вопросы теории интеллектуальных машин. — Киев: ИК АН УССР, 1975.

[3] Об автоматическом переводе математических текстов, записанных на языке TL / К.П.Вершинин, А.П.Жежерун // Автоматическая обработка математических текстов. — Киев: ИК АН УССР, 1980.

Надійшла до редколегії 23 квітня 1999 р.