УДК (038):[655.5:681.3]

І. С. Кравчук

Харківський національний університет імені В. Н. Каразіна

Семантизація синтаксичних зв'язків і її місце в різних системах автоматичного опрацювання тексту

Кравчук І. С. Семантизація синтаксичних зв'язків і її місце в різних системах автоматичного опрацювання тексту. Стаття присвячена алгоритмічному розпізнаванню типу синтаксичних зв'язків між компонентами словосполучень при автоматичному синтаксичному аналізі. Розглядаються різні типи інтерпретації синтаксичних зв'язків: семантичні, трансформаційні, перекладні й тощо. Характеризуються різні сфери використання кожного з них. Запропоновано узагальнення усіх цих типів за допомогою відповідного алфавітного оператора. Описано результати побудови абстрактного автомата для реалізації вказаного алфавітного оператора.

Ключові слова: синтаксичні зв'язки, тип синтаксичних зв'язків, семантизація зв'язків, алфавітний оператор, абстрактний автомат.

Кравчук И. С. Семантизация синтаксических связей и её место в разных системах автоматической обработки текста. Статья посвящена алгоритмическому распознаванию типа синтаксических связей между компонентами словосочетаний при автоматическом синтаксическом анализе. Рассматриваются различные типы интерпретации синтаксических связей: семантический, трансформационный, переводной и другие. Характеризуются разные сферы использования каждого из них. Предложено обобщение всех этих типов с помощью соответствующего алфавитного оператора. Описаны результаты построения абстрактного автомата для реализации указанного алфавитного оператора.

Ключевые слова: синтаксические связи, тип синтаксических связей, семантизация связей, алфавитный оператор, абстрактный автомат.

Kravchuk I. S. Syntactic relations semantisation and its position in a variety of automatic text processing. This paper is devoted to an algorithmic type recognition of syntactic relations between word-combination components during an automatic syntactic analysis. It also examines various types of syntactic relations interpretations: semantic, transformational, translational and others; as well as it characterizes different spheres of their particular usage. Moreover, the paper gives a basis for generalization of all these types through a corresponding alphabetic operator and describes the results in creation of an abstract automaton to realize the given alphabetic operator.

Key words: syntactic relation, a syntactic relation type, semantisation of connections, an alphabetic operator, abstract automaton.

У галузі автоматичного опрацювання текстів існує багато задач, які включають наступну часткову проблему. Між двома словоформами речення встановлено синтаксичний зв'язок — треба визначити тип цього зв'язку. Наприклад, два словосполучення *смеяться над недостатками* і *лететь над облаками* однакові з формально-синтаксичної точки зору, але відрізняються за смисловими відношеннями. У подальшому дану задачу ми будемо називати семантизацією (або інтерпретацією) синтаксичних зв'язків.

У залежності від цілей опрацювання тексту тип синтаксичних зв'язків може характеризувати смислові (змістові) відношення між компонентами словосполучення, трансформаційний потенціал словосполучення або його перекладні (міжмовні) можливості. Таким чином, семантизація може бути перекладною, трансформаційною і смисловою.

Перший вид семантизації пов'язаний з переходом (трансфером) від алгоритму аналізу до алгоритму синтезу. Найчастіше як одиниці перекладу при цьому використовуються конфігурації. Більшість конфігурацій виявляється багатозначними. Тому виникає задача вибору одного з перекладних відповідників, інакше кажучи, розрізнення конфігураційної омонімії. Вибір відповідника мусить враховувати лексико-синтаксичні особливості одного з компонентів конфігурації або одночасно двох. Різновидом задачі розв'язання омонімії конфігурацій є переклад прийменників. Дійсно, можна дійти згоди, у відповідністю з якою при відображенні конкретних словосполучень у конфігурацію прийменник не замінювався б символом класу і був би

такою ж розрізняльною ознакою конфігурації, як індекси відмінків.

Автоматичний переклад може бути здійснено й інакше. Конфігурація вхідної мови інтерпретується не за допомогою конфігурацій перекладаючої мови, а засобами проміжної мовної системи — семантичної мови-посередника. У цьому випадку синтаксична структура речення відображається не просто стрілками залежностей, а нумерованими (змістовими) стрілками.

По суті аналогічна задача виникає і при так званому операційному аналізі [1]. Метою його є подання синтаксичної структури у вигляді кореляцій. Кореляція являє собою сполучення двох мовних елементів (корелятів) за допомогою третього (корелятора). Кореляція для наочності зображується прямокутником виду (1 і 2 — кореляти, 3 — корелятор):

3	
1	2

Візьмімо, наприклад, словосполучення *гуляти в лісі*. За правилами кореляційного аналізу це словосполучення мусять бути зображено у вигляді наступної кореляційної форми:

30	
гуляти	*ліс*

Індекс 30 розшифровується як "місцезнаходження в межах даного простору".

Ще одна можливість інтерпретації синтаксичних зв'язків відкривається при автоматизації трансформаційного аналізу. Інтерпретація в даному випадку полягає у встановленні сукупності (іноді пустої) трансформацій, яким можна піддати дане словосполучення. Класи словосполучень, отримані при смисловій інтерпретації і при трансформаційній, часто не збігаються. Наприклад, словосполучення *бросать камни* и *двигать стулья* виражають однакові смислові (об'єктні) відношення, але відрізняються за трансформаційними можливостями [10]. Трапляються випадки, коли, незважаючи на збіг трансформаційного потенціалу, словосполучення явно відрізняються за типом смислових відношень між компонентами. Так, словосполучення *принять душ, принять гостя, принять микстуру* за допомогою трансформацій розрізненими бути не можуть. Таким чином, у зв'язку з відсутністю взаємнооднозначної відповідності між формальними і семантичними відношеннями між компонентами словосполучень смислову і трансформаційну інтерпретацію слід розглядати як дві самостійні задачі.

Трансформаційна інтерпретація є моделлю здатності мовця утворювати певні перифрази. Крім того, її можна використовувати й в одній схемі перекладу [9]. Відповідно до цієї схеми процес перекладу мусить відбуватися наступним чином. Спочатку речення X_i мови L_1 заміняється реченням Xj тієї ж мови L_1. При цьому форма речення X_j обирається так, щоб перехід від X_j до речення Y_j мови L_2 здійснювався найпростішим чином (в ідеалі шляхом елементарного перекодування). Після цього речення Y_j заміняється реченням Y_i тієї ж мови L_2. Як бачимо, значну частину процесу перекладу становлять внутрішньомовні перетворення, здійснювані за допомогою трансформацій. Виявити ж можливість тих або інших трансформацій для конкретного речення можна шляхом і трансформаційної інтерпретації синтаксичних зв'язків. Тут може виникнути питання відносно того, яким чином співвідносяться між собою трансформації словосполучень і трансформація цілих речень. Спеціальний розгляд цього питання дозволяє зробити висновок, що «трансформация всякого многочленного комплекса представляет собой сумму трансформаций бинарных комплексов» [13:52].

При перекладі речення на проміжну мову-посередник виникає необхідність ще в одному способі інтерпретації синтаксичних зв'язків [6, 7]. До висловлювання на мові смислу пред'являється вимога, щоб воно було «потенційно повним» і являло б собою «у розгорнутому вигляді усі скорочення й еліпсис, якими рясніють висловлювання у природній мові» [8]. Існує два види еліпсису: граматичні й семантичні. Для перших існують формальні ознаки, які дозволяють заповнити еліпсис. Для других формальних ознак немає. Приклад граматичного еліпсису: «Я пішов направо, він — наліво». Формально-граматична структура речення після коми вказує на наявність еліпса. У такому ж випадку як «дорога в ліс», для якого повний варіант — «дорога, що веде в ліс», немає формальних показників ні наявності еліпсису, ні його особливостей. Характер еліпсисів у подібних випадках визначається механізмом взаємодії значень компонентів у межах словосполучення.

Описаного виду інтерпретацію можна розглядати як різновид трансформаційної інтерпретації. Різниця полягає в тому, що у другому випадку не допускається вставлення нових лексем, у першому випадку — вона припускається. Усунення семантичного еліпсису в межах словосполучення може виявитися важливим ще в одній схемі перекладу.

Відповідно до цієї схеми переклад мусить відбуватися у такий спосіб [5, 12]. Спочатку

кожне речення мови, з якої здійснюється переклад, розбивається на кілька примітивних фраз. Під примітивною фразою розуміється *n*-місний предикат і *n* його аргументів. Наприклад, у фразі *Рівнобічний трикутник належить до множини правильних багатокутників* слова *трикутник, множина, багатокутник* є одномісними предикатами, слово *належить* виступає в ролі двомісного предиката. Відповідно до цього у вхідній фразі виділяються такі примітивні фрази: *належати* (*трикутник, множина*), *трикутник* (*рівнобічний*), *множина* (*багатокутники*), *багатокутник* (*правильний*).

Після виокремлення усіх примітивних фраз здійснюється їх переклад іншою мовою. Припускається, що з огляду на елементарність смислових зв'язків, а також через те, що мови менше розрізняються на рівні примітивних фраз, ніж на рівні непримітивних, такий переклад виявиться значно простішим. Заключний етап — об'єднання примітивних фраз у непримітивні.

Ця схема дещо нагадує конфігураційний трансфер від аналізу до синтезу тексту. Існують однак істотні відмінності. При конфігураційному аналізі і одиницями аналізу, і одиницями перекладу є суто граматичні одиниці, до того ж стандартної довжини (найчастіше бінарні); у другому випадку одиницями аналізу і відповідно перекладу є лексико-граматичні одиниці змінної довжини, яка залежить від числа ***n*** місць даного предиката.

Друга відмінність полягає в необхідності такої інтерпретації синтаксичних зв'язків, при якій відбувається усунення семантичного еліпсису. Це зумовлено тим, що речення мусить бути розбито саме на примітивні фрази. Елементарні ж синтаксичні одиниці — словосполучення — можуть являти собою і непримітивні фрази. Це пояснюється способами поєднання примітивних фраз у речення. Поряд із такими способами, як використання логічних зв'язок і кванторів (*і, або, якщо; усі, іноді, ...*), поєднання (одне слово входить одночасно до кількох примітивних фраз), існує і такий спосіб, як злиття. Він полягає у тому, що спільне слово двох примітивних фраз вилучається. Наприклад: *сидіти* (*людина*) + *за* (*сидіти, кермо*) → *за* (*людина, кермо*). Очевидно, для того щоб у подібних випадках виокремити примітивні фрази, необхідно здійснити вставку відсутнього елемента.

Потреба в семантизації зв'язку виникає не тільки у словосполученні, але й у межах слова. Це цілком природне узагальнення проблеми, яке гармонує з розумінням синтагми в структурній лінгвістиці. Таким чином, існує можливість описати в єдиних термінах деякі відношення між морфемами в межах слова й між словами в межах словосполучення. Аналогічну спробу зробив Ч. Хокетт, включивши до ідіом як сполучення морфем, так і сполучення слів.

Усі розглянуті види семантичної інтерпретації, незважаючи на змістову і формальну відмінність, виявляють функціональну схожість. Для того, щоб зробити їх схожість очевидною, треба показати, що вони звідні до однакової більш абстрактної задачі. Для цього найбільш зручним виявляється алфавітне зображення розглядуваних задач у вигляді так званого алфавітного оператора, під яким розуміється функція, яка встановлює відповідність між словами в одному й тому ж або в різних алфавітах.

Будь-який процес у принципі може бути зведений до певного алфавітного оператора. Побудова алфавітних операторів виправдовується такими міркуваннями. Перш за все, алфавітний оператор дозволяє виявити структурну схожість зовні різних явищ. Завдяки цьому відкривається можливість в однаковий спосіб розв'язувати більш широке коло задач, ніж це було можливо до алфавітного зображення процесу. Крім того, оператор фіксує лише релевантні риси об'єкту і, отже, частково спрощує його. Поряд із цим, оскільки функціонування оператора визначається попередньою угодою, то з'являється можливість повнішого дослідження функціонування вихідного об'єкта в різних логічно уявних ситуаціях. Усе це робить алфавітний оператор знаряддям для дослідження і «підпорою для мислення». І нарешті, алфавітне зображення процесів дозволяє звести їх опис до добре вивчених математичних задач, зокрема, до задач абстрактної теорії автоматів і скористатися вже відомими методиками.

Алфавітний оператор являє собою множину рядків відповідності між словами в певних алфавітах. Такі множини можуть бути заданими або переліком у вигляді таблиці відповідностей, або описом у вигляді алгоритму. У більшості випадків, які стосуються природних мов, задання операторів переліком є неможливим. Для того, щоб описати функціонування якої-небудь мовної системи, необхідно спочатку представити її функціонування у вигляді алфавітного оператора, а потім знайти алгоритмічний спосіб його задання.

Покажемо тепер, як звести процес інтерпретації синтаксичних зв'язків до алфавітного оператора. Будь-яке словосполучення можна

представити у вигляді формули бінарного відношення: ***xRy*** або ***R(x,y)***, де символи в дужках відповідають основам повнозначних слів, а ***R*** — показнику синтаксичних відношень, який може бути виражений різними граматичними способами. У деяких випадках показник синтаксичних відношень може складатися з кількох компонентів. З метою уніфікації алфавітних зображень складний показник будемо розглядати як простий, неподільний. Літера ***x*** у виразі ***xRy*** означає основу головного компонента словосполучення, а ***y*** — основу залежного. Як бачимо, відмінність літер відображає відмінність не тільки самих компонентів, але й їх функцій. Тому порядок символів у виразі є несуттєвим й іноді не збігається з порядком відповідних їм компонентів словосполучення. Інакше кажучи, вираз ***xRy*** слід трактувати не як ланцюжок, а як комплекс нелінійних елементів.

Кожному словосполученню можна поставити у відповідність певний тип відношень між компонентами. Перенумеруймо ці відношення і будемо називати їх номерами значень. Конкретне словосполучення і його номер значення утворюють відповідність виду:

$$xRy \rightarrow i.$$

Множина таких відповідностей утворює алфавітний оператор. Домовимось у подальшому будувати алфавітні оператори окремо для кожного фіксованого ***R***. У цьому випадку сам символ корелятора виявляється зайвим, і для алфавітного представлення стає достатнім двох символів: головного і залежного корелятів.

Кожний лінгвістичний опис мусить характеризуватись «пояснювальною силою» [11], тобто придатністю для аналізу додаткового мовного матеріалу, не використовуваного при побудові даної лінгвістичної моделі. Тому алфавітний оператор слід задати не вигляді таблиці відповідностей, а у вигляді системи правил. Такі правила розглядаються в теорії формальних граматик [3, 11] і в абстрактній теорії автоматів [2, 4]. У першій з них описуються функції породження формальних мов, у другій — функції перетворення алфавітних ланцюжків (алфавітних операторів). Очевидно, що інтерпретація синтаксичних зв'язків відноситься саме до функцій перетворення, а способом алгоритмічного завдання таких функцій можуть бути абстрактні автомати Мілі й Мура [2].

Застосування алгоритмів синтезу автоматів, що реалізують алфавітні оператори інтерпретації синтаксичних зв'язків, свідчить, що такі автомати у режимі екзамену не виявляють здатності до екстраполяції, тобто правильно опрацьовувати нові дані, що не використовувались при синтезі цих автоматів.

Ці вади можна усунути, якщо у вхідному алфавітному операторі використовувати не індивідні словоформи, а класи словоформ, наприклад, семантичні, дистрибутивні або валентностні класи. Але це мусить бути темою окремого дослідження.

Література

1. Глазерсфельд Э. «Мультистор» — система корреляционного анализа для английского языка / Э. Глазерсфельд // Автоматический перевод. — М. : Прогресс, 1971. — С. 281–318.

2. Глушков В.М. Введение в кибернетику / В. М. Глушков. — К.: Изд. АН УССР, 1964. — 324 с.

3. Гросс М., Лантен А. Теория формальных грамматик / М. Гросс, А. Лантен. — М. : Мир, 1971. — 294 с.

4. Карпов Ю.Г. Теория автоматов / Ю.Г. Карпов. — СПб.: Питер, 2002. — 206 с.

5. Леонтьева Н.Н. О представлении смысла текста набором смысловых ядерных предложений / Н.Н. Леонтьева // Тезисы докладов. — Ереван, 1967. — С. 81–84.

6. Леонтьева Н.Н. Об уровнях и оценке текстовой смысловой неполноты / Н.Н. Леонтьева // Труды международной конференции «ДИАЛОГ 2007». — М., 2007. — С. 123–131.

7. Леонтьева Н.Н., Никитина С.Е. Семантика предлогов с точки зрения автоматической обработки текста / Н.Н. Леонтьева, С.Е. Никитина // Международный семинар по машинному переводу. — М., 1975. — С. 74–75.

8. Мартемьянов Ю.С. К описанию смысла слов для целей вывода / Ю.С. Мартемьянов // Машинный перевод и прикладная лингвистика. — М., 1969. — № 12. — С. 70–83.

9. Ревзин И.И., Розенцвейг В.Ю. Основы общего и машинного перевода / И.И. Ревзин, В.Ю. Розенцвейг. — М.: Высшая школа, 1964. — 243 с.

10. Уорс Д.С. Трансформационный анализ конструкций с творительным падежом в русском языке / Д.С. Уорс // Новое в лингвистике. — М.:ИЛ, 1962. — Вып. II. — С. 637–683.

11. Хомский Н., Миллер Дж. Введение в формальный анализ естественных языков / Н. Хомский, Дж. Миллер. — М.: Едиториал УРСС, 2003. — 64 с.

12. Чулик К. Использование абстрактной семантики и теории графов в многоязычных переводных словарях / К. Чулик // Проблемы кибернетики. —М.: Физматгиз, 1965. — Вып. 3. — С. 221– 232.

13. Шаумян С.К., Соболева П.А. Аппликативная порождающая модель и исчисление трансформаций в русском языке / С.К. Шаумян, П.А. Соболева. — М.: Наука, 1963. — 125 с.