УДК 004.93:81

ЗАСТОСУВАННЯ ПОРОДЖУВАЛЬНИХ ГРАМАТИК ДЛЯ ГЕНЕРУВАННЯ РЕЧЕНЬ УКРАЇНСЬКОЮ МОВОЮ

Т.В. Шестакевич
Асистент*
Контактний тел.: (032) 292-78-65
E-mail: victana@bk.ru
В.А. Висоцька
Ассистент*
Контактний тел.: (032) 258-25-38
E-mail: victana@bk.ru
*Кафедра інформаційних систем та мереж
Національний університет "Львівська політехніка"
вул. С. Бандери, 12, м. Львів, Україна, 79013

У статті подано застосування породжувальних граматик у лінгвістичному моделюванні
Ключові слова: породжувальні граматики, синтаксичний аналіз

В статье представлено применение порождающих грамматик в лингвистическом моделировании
Ключевые слова: порождающие грамматики, синтаксический анализ

This paper presents a generative grammar application in linguistic modeling
Keywords: generative grammar, syntactic analysis

Вступ

Стрімкий розвиток Інтернету активізував створення розмаїтих лінгвістичних ресурсів, потреба в автоматизації процесів аналізу та синтезу природномовних текстів зумовила появу відповідних лінгвістичних моделей процесів їх обробки. З часом украй необхідним став розвиток багатьох мовознавчих дисциплін саме для потреб інформаційних наук.

Аналіз літературних даних і постановка проблеми

Традиційно аналіз природномовних текстів складається з трьох послідовних процесів – морфологічного, синтаксичного та семантичного аналізу. Для кожного з цих етапів були створені відповідні моделі та алгоритми. Так, теорія породжувальних граматик, початок якої був закладений у роботах американського лінгвіста Н. Хомскі [4, 5, 8-11, 13, 14], є ефективним інструментом лінгвістичного моделювання на синтаксичному рівні мови. Н.Хомські використав прийом формального аналізу граматичної структури фраз, який дає змогу виділити синтаксичні структури (складові), які є основною схемою фрази, незалежно від її значення. Ідеї Н.Хомські розвивав, серед інших, радянський лінгвіст А.В. Гладкий, який застосовував поняття дерев залежності та систем складових для моделювання синтаксичного рівня мови [1, 4, 5, 7, 11]. Він також запропонував спосіб моделювання синтаксису за допомогою синтаксичних груп, що виділяють складові словосполучень як одиниці побудови дерева залежностей, – це дало змогу об'єднати переваги методу безпосередніх складових і дерев залежностей [2, 11].

Напрацювання Н.Хомські, А.В.Гладкого застосовні до розроблення інформаційно-пошукових систем, систем машинного перекладу, анотування текстів, морфологічного, синтаксичного, семантичного аналізу текстів, навчально-дидактичні систем ті інших засобів опрацювання природної мови.

Мета і задачі дослідження

В межах статті покажемо способи застосування апарату породжувальних граматик до моделювання синтаксису речень для української мови. Для цього розберемо синтаксичну структуру речень, продемонструємо особливості процесу синтезу речень українською мовою. Розглянемо, як впливають норми та правила мови на хід побудови граматик.

Основний матеріал

Формальна породжувальна граматика G – це четвірка G = (V, T, S, P), де V – *алфавіт*, T – *термінальні символи*, S – *початковий символ* (S∈V), P – скінченна множина *продукцій* вигляду $\xi \to \eta$, де ξ та η – ланцюжки над V. Множина *нетермінальних символів* N=V\T. Вважатимемо термінальні символи словоформами (деякої природної мови), нетермінальні символи – синтаксичними категоріями, а термінальні ланцюжки, що виводяться, – правильними реченнями даної мови. Тоді виведення речення інтерпретується як його синтаксична структура, подана в термінах безпосередній складових. Граматики класифікують за типами продукцій та виділяють довільний, контекстно-залежний, контекстно-вільний та регулярний тип.

Розглянемо процес виведення речення в українській мові. Українській мові властивий вільний порядок слів у реченні, що, втім, не заперечує існування усталеного порядку розташування окремих мовних елементів. Для простого повного речення з прямим порядком слів структурну схему вважатимемо фіксованою, основними синтаксичними категоріями такого речення будуть іменна та дієслівна групи [12]. Розглянемо породжувальну граматику довільного типу для моделювання синтаксису речення українською мовою вказаної структурної схеми. Алфавітом є іменникова група, дієслівна групи та їх складові (нетермінальні символи), а також лексичний запас української мови (відповідні термінальні символи). Усі можливі перетворення термінальних символів у нетермінальні складають множину правил. Під час виведення отримують незліченну кількість термінальних ланцюжків (речень українською відповідної структурної схеми), тому така граматика буде необмеженою, і через свою складність застосування не матиме.

Для введення контекстно-залежної граматики накладемо такі обмеження на структуру речення. Спираючись на правила побудови речень української мови з прямим порядком слів (наприклад, прикметник стоїть у препозиції до іменника [3], розглянемо іменну групу структурної схеми, де прикметник та іменник в іменній групі узгоджуються між собою за відмінком, числом та родом:

Прикметник Іменник (1)

Розглядатимемо дієслівну групу такої структурної схеми:

Дієслово Прислівникова група (2)

Розглядатимемо *прислівникову* групу таку структурну схему (у квадратних дужках зазначено необов'язкові елементи, у фігурних – елементи, які можуть повторюватись):

{Прислівник}[{Прикметник}] (3)

З огляду на граматичні характеристики дієслова в українській мові, узгодження між іменною та дієслівною групою відбувається за числом, родом та особою. Складові іменної групи (після «/» тут і надалі вказано використовувані позначення)/ $\widetilde{N}$: прикметник/A, іменник/N. Граматичні категорії іменної групи та її складових: число/ЧЛ: однина/од, множина/мн; рід/РД: чоловічий/ч, жіночий/ж, середній/с; відмінок/ВД: називний/н, родовий/р, давальний/д, знахідний/з, орудний/о, місцевий/м, кличний/к; особа/ОС: 1-ша/1, 2-га/2, 3-тя/3. Відповідно, скороченим позначеннями іменної групи буде $\widetilde{N}_{\text{РД,ЧЛ,ВД,ОС}}$, а її складових – $A_{\text{РД,ЧЛ,ВД}}$, $N_{\text{РД,ЧЛ,ВД,ОС}}$. За потреби акцентувати на використанні різних значень граматичних категорій, наприклад, категорії числа для двох іменних груп, вживатимемо такі позначки: $\widetilde{N}_{\text{РД,ЧЛ',ВД,ОС}}$, $\widetilde{N}_{\text{РД,ЧЛ'',ВД,ОС}}$. Складові дієслівної групи/ $\widetilde{R}$: дієслово/R та в межах прислівникової групи/ $\widetilde{B}$ прикметник (описаний вище) та прислівник/B. Граматичні категорії дієслівної групи та її складових: число/ЧЛ: однина/од, множина/мн; рід/РД: чоловічий/ч, жіночий/ж, середній/с; особа/ОС: 1-ша/1, 2-га/2, 3-тя/3; час/ЧС: теперішній/тп, минулий/мн, майбутній/мб. Відповідно, скороченим позначеннями дієслівної групи буде $\widetilde{R}_{\text{РД,ЧЛ,ЧС,ОС}}$, дієслова – $R_{\text{РД,ЧЛ,ЧС,ОС}}$, прислівникової групи – $\widetilde{B}_{\text{РД,ЧЛ,ВД}}$. Універсальною складовою речення введемо сполучник/Q.

Спосіб задання контекстно-залежної граматики, що виводить речення введеної структурної схеми (з урахуванням закономірностей української мови) приведемо на прикладі речення *Їхній одяг виглядав однаково вологим і пошарпаним*, запозиченого із [6]. Розглянемо граматику G=(V, T, S, P). Алфавіт (позначення синтаксичних категорій подамо без індексів – для зручності) V=(S, $\widetilde{N}$, $\widetilde{R}$, A, N, R, B, $\widetilde{B}$, Q, *їхній, одяг, виглядати, однаково, вологий, і, пошарпаний*), T=(*він, одяг, виглядати, однаково, вологий, і, пошарпаний*), S – початковий символ, множину правил P подамо у вигляді табл. 1.

Таблиця 1

Правила формування речення українською мовою

№	Розгортання	Правила
I.	Структури	$S\rightarrow\#\,\widetilde{N}_{\text{РД,ЧЛ,ВД,ОС}}\,\widetilde{R}_{\text{ЧЛ,ЧС,ОС}}\#$.
II.	Іменної групи	$\widetilde{N}_{\text{РД,ЧЛ,ВД,З}}\rightarrow A_{\text{РД,ЧЛ,ВД}}N_{\text{РД,ЧЛ,ВД,З}}$.
III.	Дієслівної групи	1) $\widetilde{R}_{\text{ЧЛ,ЧС,ОС}}\rightarrow R_{\text{ЧЛ,ЧС,ОС}}\,\widetilde{B}_{\text{РД,ЧЛ,ВД}}$; 2) $\widetilde{B}\rightarrow BA_{\text{РД,ЧЛ,ВД}}$; 3) $BA_{\text{РД,ЧЛ,ВД}}\rightarrow BA_{\text{РД,ЧЛ,ВД}}QA_{\text{РД,ЧЛ,ВД}}$. Це правило вводить контекстні обмеження: якщо прислівникова група містить кілька прикметників, то їх моделює прислівник групи, і ці прикметники вживаються з однаковими граматичними характеристиками.
IV.	Синтаксичних категорій словоформами	1)$N_{\text{ч,ЧЛ,ВД}}\rightarrow$одяг$_{\text{ЧЛ,ВД}}$, ...; 2) $R_{\text{ЧЛ,ЧС,ОС}}\rightarrow$ $\rightarrow$виглядати$_{\text{ЧЛ,ЧС,ОС}}$,...; 3) Q$\rightarrow$і,...; 4) $A_{\text{РД,ЧЛ,ВД}}\rightarrow$вологий$_{\text{РД,ЧЛ,ВД}}$, пошарпаний$_{\text{РД,ЧЛ,ВД}}$,...; 5) B$\rightarrow$однаково; 6) $A_{\text{РД,мн,ВД}}\rightarrow$їхній$_{\text{РД,ВД}}$.

Виведення речення заданої структурної схеми подано далі.

1. S
2. (I) #$\tilde{N}_{\text{ч,од,н,З}}\,\tilde{R}_{\text{од,мн,З}}$#
3. (II) #$A_{\text{ч,мн,н}}N_{\text{ч,од,н}}\,\tilde{R}_{\text{од,мн,З}}$#
4. (III.1) #$A_{\text{ч,мн,н}}N_{\text{ч,од,н}}R_{\text{од,мн}},3\,\tilde{B}_{\text{ч,од,о}}$#
5. (III.2) #$A_{\text{ч,мн,н}}N_{\text{ч,од,н}}R_{\text{од,мн,З}}BA_{\text{ч,од,о}}$#
6. (III.3) #$A_{\text{ч,мн,н}}N_{\text{ч,од,н}}R_{\text{од,мн,З}}BA_{\text{ч,од,о}}QA_{\text{ч,од,о}}$#
7. (IV.6) #Їхній $N_{\text{ч,од,н}}R_{\text{од,мн,З}}BA_{\text{ч,од,о}}QA_{\text{ч,од,о}}$#
8. (IV.1) #Їхній одяг $R_{\text{од,мн,З}}BA_{\text{ч,од,о}}QA_{\text{ч,од,о}}$#
9. (IV.2) #Їхній одяг виглядав $BA_{\text{ч,од,о}}QA_{\text{ч,од,о}}$#
10. (IV.5) #Їхній одяг виглядав однаково $A_{\text{ч,од,о}}$ $QA_{\text{ч,од,о}}$#
11. (IV.4) #Їхній одяг виглядав однаково вологим $QA_{\text{ч,од,о}}$#
12. (IV.3) #Їхній одяг виглядав однаково вологим і $A_{\text{ч,од,о}}$#
13. (IV.4) #Їхній одяг виглядав однаково вологим і пошарпаним#

Приклад виведення для генерування речення українською подано на рис. 1.

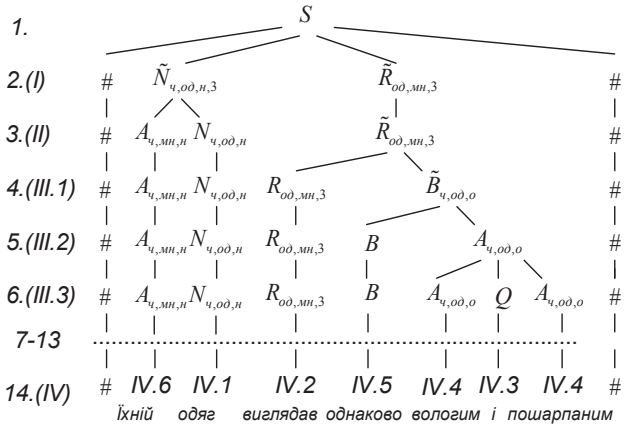

Рис. 1. Виведення в контекстно-залежній граматиці G

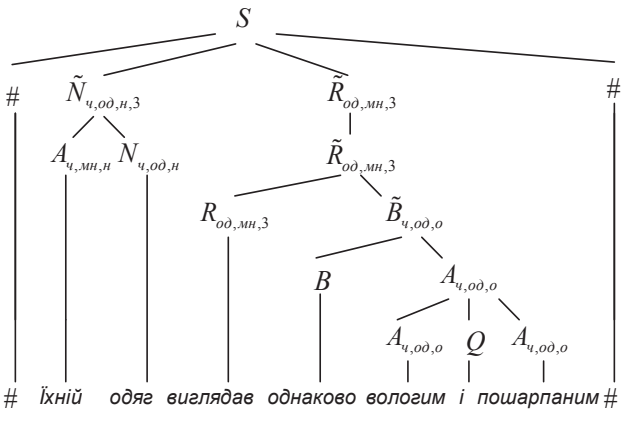

Рис. 2. Дерево складових для контекстно-залежної граматики G

Кожен крок виведення полягає або в розгортанні одного з символів попереднього ланцюжка (так, при переході від ланцюжка 4 до ланцюжка 5 символ $A_{ч,од,н}$ розгортається в три символи – $A_{ч,од,н}$ Q $A_{ч,од,н}$), або в заміні його іншим символом, усі інші символи переписуються без змін. Якщо розгортані, замінювані або переписувані символи з'єднати лініями безпосередньо з символами, які виходять в результаті розгортання, заміни або переписування, отримаємо дерево складових, або синтаксичну структуру (рис. 2). Прикладом правила з використанням контексту в граматиці G є правило III.3. Відмова від контексту робить структуру граматик простішою та полегшує їх дослідження. Особливими є випадки, коли контекст змістовно необхідний, але формально його враховують за допомогою контекстно-вільних правил через введення в граматику нових категорій [4, 5].

Висновки

Апарат породжувальних граматик, запропонований Н.Хомські, моделює процеси на синтаксичному рівні мови – виділені структурні елементи речення дають змогу описувати синтаксичні конструкції незалежно від їх змісту.

У статті показано особливості процесу синтезу речення українською мовою із застосуванням породжувальних граматик, розглянуто вплив норм та правил мови на хід побудови граматик. Застосування породжувальних граматик має широкі можливості у розробленні автоматизованих систем опрацювання текстів, лінгвістичного забезпечення комп'ютерних лінгвістичних систем тощо.

Література

1. Анісімов А.В. Алгоритмічна модель асоціативно-семантичного контекстного аналізу текстів природною мовою / А.В. Анісімов, О.О. Марченко, А.О. Никоненко // Пробл. програмув. – 2008. – N 2-3. – С. 379-384.
2. Апресян Ю. Д. Непосредственно составляющих метод // Лингв. энцикл. словарь / Под ред. В. Н. Ярцевой. – М. : Сов. энциклопедия, 1990.
3. Багмут А.Й. Порядок слів // Українська мова: Енцикл. – 3-тє вид. / А.Й. Багмут. – К. : В-во "Укр. енциклопедія" ім. М.П. Бажана, 2007. - С.507-508.
4. Гладкий А.В. Элементы математической лингвистики / А.В. Гладкий, И. А. Мельчук. – М. : Наука, 1969. – 192 с.
5. Гладкий А. В. Формальные грамматики и языки / А.В. Гладкий. – М. : Наука, 1973. – 368 с.
6. Кінг С. Зона покриття / С. Кінг. – Харків : Книжковий Клуб "Клуб Сімейного Дозвілля", 2008. – 432 с
7. Марченко О.О. Алгоритми семантичного аналізу природномовних текстів : автореф. дис. на здобуття наук. ступеня канд. фіз.-мат. наук: спец. 01.05.01 // О.О. Марченко. – Київ, 2005.
8. Хомски Н. Формальный анализ естественных языков / Н. Хомски, Дж. Миллер // Кибернетический сборник. – М. : Мир, 1965. – № 1. – С. 231-290.
9. Хомски Н. Язык и мышление / Н. Хомски // Публикации ОСиПЛ. Серия монографий. – М.: Изд-во Московского университета, 1972. – № 2. – 122 с.
10. Хомски Н. Синтаксические структуры / Н. Хомски // Сборник «Новое в лингвистике». – М.: ИЛ, 1962. – № 2. – С. 412-527.
11. Шаров С.А. Средства компьютерного представления лингвистической информации [Електронний ресурс] // Режим доступу : http://www.ksu.ru/eng/science/ittc/vol000/002/.
12. Шульжук К. Синтаксис української мови: Підручник / К. Шульжук. – К. : Академія, 2004. – 397 с.
13. Chomsky N. Formal properties of grammars / N. Chomsky. // Handbook of Mathemati-Mathematical Psychology, 2, ch. 12, Wiley, 1963. – P. 323-418.
14. Chomsky N. Introduction to the formal analysis of natural languages / N. Chomsky, G. A. Miller // Handbook of Math. Psyc.2, Wiley, 1963. – P. 269-322.