УДК 81:004.93

В. А. Висоцька
Національний університет “Львівська політехніка”,
кафедра інформаційних систем та мереж

КОНЦЕПТУАЛЬНА МОДЕЛЬ ПРОЦЕСУ ФОРМУВАННЯ СЕМАНТИКИ РЕЧЕННЯ ПРИРОДНОЮ МОВОЮ

Наведено застосування породжувальних граматик у лінгвістичному моделюванні. Опис моделювання синтаксису речення застосовують для автоматизації процесів аналізу та синтезу природномовних текстів. Запропоновано методи аналізу контенту для інтернет-газети. Модель описує процеси опрацювання інформаційних ресурсів у системах аналізу контенту та спрощує технологію автоматизації управління контентом. Проаналізовано основні проблеми синтаксичного та семантичного аналізу контенту та функціональних сервісів управління контентом. Контент у вигляді статей є основою інтернет-газети, за яким користувач шукає необхідну йому інформацію. За допомогою методу аналізу контенту власник системи може визначити достовірність та оперативність інформації, розміщеної в інтернет-газеті на інформаційному ресурсі. Завдяки цьому можна визначати популярність газети та здійснювати певні дії для збільшення кількості користувачів, підвищення популярності тощо. Розроблено загальні рекомендації з проектування архітектури систем аналізу контенту, які відрізняються від інших більшою деталізацією етапів та наявністю модулів опрацювання інформаційних ресурсів, що дають змогу ефективно та просто опрацьовувати інформаційні ресурси на рівні розробника систем.

Ключові слова: породжувальні граматики, структурна схема речення, комп’ютерна лінгвістична система, контент, аналіз контенту, інформаційний ресурс, система управління контентом.

This paper presents the generative grammar application in linguistic modelling. Description of syntax sentence modelling is applied to automate the processes of analysis and synthesis of texts in natural language. This article suggests methods of content analysis for online newspaper. The model describes the processing of information resources systems of content analysis and simplifies the technology of content management system automation. General problems of syntactical and semantic content analysis and functional services of content management system are analysed. Content as articles is the base of online newspaper due to which the user is looking for the necessary information. Thanks to content analysis, the system owner can determine the reliability and efficiency of the information contained in the articles of online newspaper. With the help of this option you can determine the popularity of the newspaper and do some actions in order to augmenter the number of users. General recommendations in architectural design of content analysis systems are developed which, however, differ from existing by more detailed stages and availability of information processing module resources, allowing an efficient and easy to handle information resources at system developer’s stage.

Key words: generative grammar, structured scheme sentences, computer linguistic system, content, content analysis, information resource, content system management.

Вступ. Загальна постановка проблеми

Теорія породжувальних граматик є ефективним інструментом лінгвістичного моделювання на синтаксичному рівні мови. Початок цієї теорії закладено у роботах лінгвіста Н. Хомські, де використано формальний аналіз граматичної структури фраз для виділення синтаксичної структури (складових) як основної схеми фрази, незалежно від її значення [38–41, 49–57]. А. Гладкий

застосовував поняття дерев залежності та систем складових для моделювання синтаксичного рівня мови [14–16]. Він запропонував спосіб моделювання синтаксису за допомогою синтаксичних груп, що виділяють складові словосполучень як одиниці побудови дерева залежностей – таке подання дає змогу об'єднати переваги методу безпосередніх складових і дерев залежностей.

Аналіз останніх досліджень і публікацій

Активний розвиток Інтернету сприяє створенню різноманітних лінгвістичних ресурсів. Необхідність в реалізації процесів аналізу та синтезу природномовних текстів зумовила появу відповідних лінгвістичних моделей процесів їх опрацювання [2–5, 8, 9, 12–19, 21–24, 26–29, 32, 33, 37–41, 43, 46, 48–59, 61–63, 68]. Необхідно розвивати мовознавчі дисципліни для потреб інформаційних наук. Інтеграційні процеси в більшості галузей сучасного життя привертають особливу увагу до розроблення та створення автоматизованих систем опрацювання багатомовної інформації. Формальна породжувальна граматика G – це четвірка $G = (V, T, S, P)$, де V – скінченна непорожня множина, *алфавіт* (*словник*); T – її підмножина, елементи якої є *термінальними* (*основними*) символами, *терміналами*; S – *початковий символ* ($S \in V$); P – скінченна множина *продукцій* (*правил перетворення*) вигляду $\xi \to \eta$, де ξ та η – ланцюжки над V. Множину $V \setminus T$ позначають N, її елементи є *нетермінальними* (*допоміжними*) символами, *не терміналами* [14–16]. Тлумачитимемо термінальні символи як словоформи (деякої природної мови), нетермінальні символи – як синтаксичні категорії, а термінальні ланцюжки, що виводяться, – як правильні речення цієї мови [12–16, 38–41, 49–57]. Тоді виведення речення природно інтерпретується як його синтаксична структура, яку подано в термінах безпосередніх складових, тобто способом, давно відомим у лінгвістиці [10, 44, 46, 47]. Напрацювання і дослідження Н. Хомські та А. Гладкого розвинули та продовжили А. Анісімов [2, 3], Ю. Апресян [4, 5], Н. Більгаєва [8], Є. Большакова, Е. Клишинський, Д. Ланде, А. Носков, О. Пескова та Є. Ягунова [9], І. Волкова та Т. Руденко [11], О. Гакман [12], А. Герасімов [13], М. Гросс та А. Лантен [17], Н. Дарчук [18], І. Демешко [19], В. Інгве [21, 65, 66], Т. Любченко [22], Б. Мартиненко [23], О. Марченко [24], Є. Падучева [26], З. Партико [27], А. Пентус та М. Пентус [28], Е. Попов [29], Г. Потапова [32], Н. Русаченко [33], В. Фомічев [37], С. Шаров [43], Ю. Щербина [46], Ю. Шрейдер [48], Ю. Бар-Гіллель та Е. Шамір [58], Д. Бобров [59], Д. Хейс [61], П. Постал [62], Л. Тесніере [63], Д. Варга [68]. Ці дослідження використовують для розроблення таких засобів опрацювання природної мови, як інформаційно-пошукові системи, системи машинного перекладу, анотування текстів, морфологічний, синтаксичний та семантичний аналіз текстів, навчально-дидактичні системи, до лінгвістичного забезпечення спеціалізованих програмних систем тощо [18, 22, 27, 43].

Формулювання мети

Розглянемо способи застосування апарату породжувальних граматик до моделювання синтаксису речень для різних природних мов, наприклад, англійської, німецької, російської та української. Для цього вивчимо синтаксичну структуру речень, продемонструємо особливості процесу синтезу речень зазначених мов. Розглянемо вплив норм та правил мови на процес побудови граматик [10, 44, 46, 47].

Аналіз отриманих наукових результатів

Породжувальна граматика G – це четвірка $G = (V, T, S, P)$, де V – скінченна непорожня множина, *алфавіт* (*словник*); T – її підмножина, елементи якої є *термінальними* (*основними*) лексичними одиницями, *терміналами*; S – *початковий символ* ($S \in V$); P – скінченна множина *продукцій* (*правил перетворення*) вигляду $\xi \to \eta$, де ξ та η – ланцюжки над V. Множину $V \setminus T$ позначають N, її елементи є *нетермінальними* (*допоміжними*) лексичними одиницями, *не терміналами* [14–16]. Граматики класифікують за типами продукцій, на які накладено певні обмеження (табл. 1) [14–16, 38–41, 49–57]. Словник V складається зі скінченної непорожньої множини лексичних одиниць [60]. Вираз над V є ланцюжком скінченної довжини лексичних одиниць із V. Порожній ланцюжок, який не містить лексичних одиниць, позначимо через Λ.

Множину всіх лексичних одиниць над V позначимо V^*. Мова над V є підмножиною V^*. Мову задають через множину всіх лексичних одиниць мови або через означення критерію, якому повинні задовольняти лексичні одиниці, щоб належати мові [14–16, 38–41, 49–57]. Ще один важливий спосіб задати мову – використати породжувальну граматику. Граматика складається з множини лексичних одиниць різного типу та множини правил або продукцій побудови виразу. Граматика має словник V, який є множиною лексичних одиниць для побудови виразів мови. Деякі лексичні одиниці словника (термінальні) не можуть замінятися іншими лексичними одиницями.

Таблиця 1

Класифікація граматик за типами продукцій

Граматика	Тип	Опис
G_0	Необмежена	Тут ξ – довільний ланцюжок, що містить хоча б один нетермінальний символ, η – довільний ланцюжок над V.
G_1	Контекстно-залежна	В множині продукцій P є продукція вигляду $\gamma\xi\delta\to\gamma\eta\delta$, $\lvert\xi\rvert \le \lvert\eta\rvert$ (але не у формі $\xi\to\eta$), то ξ можна замінити на ν лише в оточенні ланцюжків $\gamma\ldots\delta$, тобто у відповідному контексті.
G_2	Контекстно-вільна	Нетермінал A у лівій частині продукції $A\to\eta$ може бути замінений ланцюжком η у довільному оточенні щоразу, коли він зустрічається, тобто незалежно від контексту.
G_3	Регулярна	Можуть бути лише продукції $A\to aB$, $A\to a$, $S\to\lambda$, де A, B – нетермінали, a – термінал, λ – порожній ланцюжок.

Термінальні лексичні одиниці є словоформами природної мови, нетермінальні лексичні одиниці – синтаксичні категорії, а термінальні ланцюжки, що виводяться, – правильні вирази цієї мови [14–16, 38–41, 49–57]. Тоді виведення виразу природно інтерпретують як його синтаксичну структуру, яку подано в термінах породжувальної граматики [14–16]. Множина виразів природною мовою має низку специфічних властивостей. Аналізуючи вирази природною мовою в теорії формальних граматик, їх розглядають як ланцюжки словоформ/морфем в ролі термінальних лексичних одиниць. Для множини виразів існує алгоритм розпізнавання, чи поданий ланцюжок є виразом певної мови. Множини, для яких існують алгоритми розпізнавання, є рекурсивними. Але для породження виразів природної мови і лише їх на граматики накладають обмеження через продукції: в продукції $A\to B$ ланцюжок B не коротший за ланцюжок A; тоді в процесі виведення ланцюжки не скорочуються.

Характерна особливість контекстно-вільних граматик G_2 – на кожному кроці виведення опрацьовується *лише один символ*, тобто жодним чином не може бути враховано наявність/відсутність або властивості різних сусідніх символів. Це може створити враження, що граматики G_2 мало придатні для опису природних мов: адже в звичайних граматиках твердження про вибір тих чи інших форм, про варіювання або розгортання тих або інших елементів вислову, як правило, формулюються саме з врахуванням контекстних умов. Так, при описі флективних форм вказують флексію, яку обирають залежно від типу основи (що є контекстом); при описі вживання українських відмінків вказують, що знахідний відмінок прямого доповнення замінюється родовим за наявності заперечення тощо; орудний суб'єкта можливий при віддієслівному іменнику лише в тому випадку, якщо при цьому іменнику є доповнення в родовому відмінку (*перегляд контенту користувачем*, але не **перегляд користувачем*) тощо [7, 10, 18–20, 32, 33, 36, 45]. Граматика G_3 практично здатна породжувати переважну більшість простих і складних речень природної мови. Тому це твердження справедливе і для довільних граматик G_2. Майже у всіх випадках, коли використання контексту на перший погляд є неминучим, фактично без нього можна обійтися. Наприклад, нехай є клас елементів X, причому в сусідстві з елементами деякого класу Y елементи X поводяться інакше, ніж в сусідстві з елементами класу Z за такими правилами:

1. $YX\to YAB$ 2. $ZX\to ZCD$ (ці правила використовують контекст).

Позначимо через X_1 елемент X в позиції після Y, а через X_2 – елемент X в позиції після Z. Тоді можна перейти до правил, що не використовують контексту:

1'. $X_1 \to AB$, 2'. $X_2 \to CD$.

Тобто вводяться більш дробові категорії елементів, що враховують їх позиції в контексті. Покажемо, як можна перейти до формулювань граматики G_2 у наведених вище прикладах звернення до контексту. Зліва поміщається потрібний фрагмент відповідної контекстно-залежної граматики, справа – еквівалентний фрагмент, що складається з контекстно-вільних правил.

а) вибір флексії відмінка залежно від типа основи [7, 10, 18–20, 32,33, 36, 45]:

$Word_{мн,род} \to O^i F_{мн,род}$ $\quad$ $Word_{мн,род} \to O^i F^i_{мн,род}$

$O^1 F_{мн,род} \to O^i ів(друз - ів)$ $\quad$ $F^1_{мн,род} \to ів$

$O^2 F_{мн,род} \to O^i ок(іграш - ок)$ $\quad$ $F^2_{мн,род} \to ок$

$O^3 F_{мн,род} \to O^i ей(дім - ей)$ $\quad$ $F^3_{мн,род} \to ей$

$O^4 F_{мн,род} \to O^i их(знайом - их)$ $\quad$ $F^4_{мн,род} \to их$

$O^5 F_{мн,род} \to O^i (машин -)$ $\quad$ $F^5_{мн,род} \to \Lambda$

…….

де $Word$ – словоформа, O^i – основа типу i $(i = 1,2,3,...)$, $F_{мн,род}$ – флексія род. відмінку мн.

б) вибір відмінка прямого доповнення залежно від наявності заперечення:

$\tilde{R} \to R^i \tilde{N}_d$ $\quad$ $\tilde{R} \to R^i \tilde{N}^1_d$

$\tilde{R} \to \neg R^i \tilde{N}_d$ $\quad$ $\tilde{R} \to \neg R^i \tilde{N}^2_d$

$XR\tilde{N}_d \to \underset{X \neq \neg x}{XR^i \tilde{N}_з}$ $\quad$ $\tilde{N}^1_d \to \tilde{N}_з$ (*студент вивчає дисципліну*)

$X\neg R\tilde{N}_d \to \underset{X \neq \neg x}{X\neg R^i \tilde{N}_p}$ $\quad$ $\tilde{N}^2_d \to \tilde{N}_p$ (*студент не вивчає дисципліну*)

де $\tilde{R}$ – група дієслова; R^i – перехідне дієслово; $\tilde{N}_d$ – пряме доповнення; $\tilde{N}$ – група іменника; $\neg$ – заперечення [7, 10, 18–20, 32, 33, 36, 45].

в) можливість орудного суб'єкта при віддієслівному іменнику залежно від наявності об'єкта (*перегляд контенту користувачем*):

$\tilde{N} \to \tilde{N}'\tilde{N}^{ob}\tilde{N}^{sb}$ $\quad$ $\tilde{N} \to \tilde{N}'\tilde{N}^{ob}\tilde{N}^{sb^1}$

$\tilde{N} \to \tilde{N}'\tilde{N}^{sb}$ $\quad$ $\tilde{N} \to \tilde{N}'\tilde{N}^{sb^2}$

$\tilde{N}^{ob}\tilde{N}^{sb} \to \tilde{N}^{ob}\tilde{N}_o$ $\quad$ $\tilde{N}^{sb^1} \to \tilde{N}'_o$

де $\tilde{N}^{ob}$ – об'єкт, $\tilde{N}^{sb}$ – суб'єкт [7, 10, 18–20, 32, 33, 36, 45]. У цих прикладах використано один і той самий формальний прийом: інформація про контекст кодується в нових категоріях. Отже, що менше використовувати контекст, то більше категорій необхідно ввести, і навпаки. Привабливість переходу до контекстно-вільних правил полягає в тому, що оцінити ступінь складності різноманітних і змістовно строкатих звернень до контексту важко, тоді як в правилах граматики G_2 мірою ступеня складності є кількість вживаних категорій.

Від контексту не можна відмовитися, тобто неможливо обійтися одним символом у лівій частині правила, якщо правило повинне забезпечувати перестановку символів. Отже, граматика G_2 не може породити мову, що містить ланцюжки, які не можуть бути побудовані без вживання перестановок. Розглянемо, наприклад, мову, що містить всілякі ланцюжки вигляду $a_1a_2a_3qa'_1a'_2a'_3$, $a_2a_1a_2a_3qa'_2a'_1a'_2a'_3$, $a_1a_3a_2a_1qa'_1a'_3a'_2a'_1$ тощо (у загальному вигляді такі ланцюжки можна записувати як xqx') і що не містить ніяких інших ланцюжків. Змістовно a_1 і a'_1, a_2 і a'_2 тощо можна розуміти

як пари елементів, певним чином узгоджених між собою. Тут йдеться не про самі символи, a_1, a_1' тощо, а про їхні відповідні входження до ланцюжків. Ця мова легко породжується граматикою, що містить правила перестановки. Для цієї граматики існує еквівалентна граматика G_1, що містить правила перестановки, яку можна подати так:

$$1.\ \left.\begin{array}{l} 1.\ I \to IA_i a_i', \\ 2.\ a_i' A_j \to A_j a_i', \\ 3.\ IA_i \to a_i I, \\ 4.\ I \to q. \end{array}\right\} i, j = 1,2,3,$$

(a_i, a_i', q – основні символи; I, A_i – допоміжні символи; I – початковий символ).

Покажемо для прикладу, як можна вивести в цій граматиці ланцюжок $a_2a_1a_1a_3qa_2'a_1'a_1'a_3'$:

1. I
2. (1) IA_3a_3'
3. (1) $IA_1a_1'A_3a_3'$
4. (1) $IA_1a_1'A_1a_1'A_3a_3'$
5. (1) $IA_2a_2'A_1a_1'A_1a_1'A_3a_3'$
6. (2) $IA_2A_1a_2'a_1'A_1a_1'A_3a_3'$
7. – 11. ..... (2; 5 разів)
12. (3) $a_2IA_1A_1A_3a_2'a_1'a_1'a_3'$
13. – 15. ..... (3; 3 рази)
16. (4) $a_2a_1a_1a_3qa_2'a_1'a_1'a_3'$

Мова $\{xqx'\}$ не може бути породжена жодною граматикою G_2. Це явище зустрічається і в природних мовах, тобто в них можливі фрагменти, що складаються з ланцюжків вигляду xqx'. У літературі описано два приклади такого роду:

1) конструкції типу: $\overset{a}{Саша}$, $\overset{b}{Софія}$, $\overset{c}{Катя}$, $\overset{d}{Данило}$, … – $\overset{a'}{спортсмен}$, $\overset{b'}{співачка}$, $\overset{c'}{художниця}$, $\overset{d}{поет}$, … *відповідно* [7, 10, 18–20, 32, 33, 36, 45]. Тут роль x $(abcd...)$ грає ланцюжок власних імен, а роль x' $(a'b'c'd'...)$ – ланцюжок професій, які мають бути погоджені з цими іменами в роді [58]; q – це тире (точніше, в'язка *є* в нульовому вираженні).

2) у індіанській мові мохавк згідно із [62] поширені речення, в яких основне доповнення дублюється інкорпорацією відповідних основ в дієслово-присудок: *Художник картино-малює пейзаж*. Крім того, будь-яке дієслово (зокрема яке містить інкорпоровані доповнення) легко субстантивується і набуває здатності виступати в ролі доповнення, зокрема, інкорпоруватися в присудок: моя дитина вподобала книгочитання (тобто «моя дитина вподобала читання книг») [7, 10, 18–20, 32, 33, 36, 45]. Цей процес теоретично може бути повторений необмежену кількість разів: він книгочитанняцікаводумає про книгочитанняцікавість («він думає про цікавість до читання книг»), тобто

$$Він\ \overbrace{книгочитанняцікавість}^{a\,b\,c} - думає\ про - \overbrace{книгочитанняцікавість}^{a'\,b'\,c'}.$$

Тут x' $(= a'b'c'd')$ – це доповнення, x $(= abcd)$ – його дублікат, інкорпорований у присудок, а q – присудок. Така конструкція є правильною лише тоді, коли інкорпорований дублікат доповнення точно відповідає доповненню за складом і порядком дотримання основ.

Враховувати ці приклади граматики G_2 недостатньо для опису будь-яких природних мов у повному обсязі. Але обидва приклади мають периферійний характер: перша конструкція, хоча і допустима, ймовірно, в будь-якій мові, украй специфічна і не належить до поширених, а друга така, що має дуже загальне значення і, мабуть, доволі поширена, відома лише в одній малопоширеній мові. Тому при всій теоретичній цінності цих прикладів ними можна нехтувати. Якщо ж на них не зважати, то граматики G_2 можна вважати в принципі достатнім засобом для опису природних мов.

Це твердження зрозуміле, але може бути строго доведено; переконання в його істинності ґрунтується на низці наступних таких міркувань.

1. Існують так звані категоріальні граматики, що належать до граматик, що розпізнають. Ці граматики розробляли і застосовували до природних мов незалежно від граматик G_2, причому прикладів їх неадекватності (за винятком двох, вказаних вище) досі наведено не було. Проте, як довів А. Гладкий [14–16], клас мов, що описуються категоріальними граматиками, збігається з класом контекстно-вільних мов.

2. Порівняно недавно для опису мов було запропоновано автомати з магазинною пам'яттю, здатні здійснювати як розпізнавання, так і породження. Н. Хомські довів, що всі мови, що опрацьовуються такими автоматами, є контекстно-вільними, і навпаки [38–41, 49–57]. Отже, ще одна формальна модель природної мови, що введена з незалежних міркувань і не стикнулася із суттєвими принциповими труднощами, еквівалентна граматиці G_2.

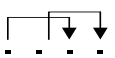
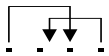
3. У межах математичної лінгвістики легко виділити клас так званих мов, що скінченно характеризуються, які інтуїтивно дуже близькі до природних мов. Всі мови, що скінченно характеризуються, є контекстно-вільними (зворотне невірно!). Це знову схиляє до думки про те, що граматики G_2 здатні породжувати природні мови.

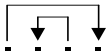
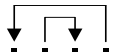
4. Нарешті, відомі алгоритми автоматичного аналізу і породження текстів природними мовами, які використовують як опис відповідних мов граматики G_2 або ж еквівалентні ним системи. На граматиках G_2 основані, наприклад, алгоритми синтаксичного аналізу для декількох мов, які розробляють у Техаському університеті [64], алгоритми, що використовують так званий метод Кока [61] і деякі інші алгоритми, згадані в роботах [6, 59, 66].

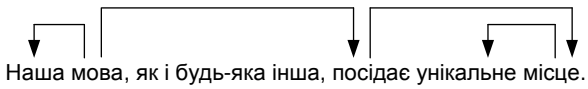
Все це дозволяє визнати граматики G_2 достатніми для природних мов. Зокрема варто зазначити, що конструкції типу $abcd...d'c'b'a'$, що не описуються граматиками G_3, легко можуть бути породжені за допомогою граматики G_2. Так, мова, яка складається точно з ланцюжків такого вигляду (складених із символів a_1, a_2, a_3, a_1', a_2', a_3'), породжується граматикою G_2, що містить лише шість правил:

$$\left.\begin{array}{l} I \to a_i I a_i' \\ I \to a_i a' \end{array}\right\} i = 1,2,3.$$

Тепер необхідно зробити два важливі зауваження.

По-перше, сказане зовсім не означає, що граматики G_2 породжують лише природні мови і/або мови, близькі до них: серед контекстно-вільних мов є і такі, які зовсім не схожі за своєю будовою на природні. По-друге, те, що граматики G_2 практично достатні для опису природних мов, зовсім не означає, що вони завжди зручні для цієї мети, тобто що вони дають змогу природно описувати будь-які конструкції природних мов. Граматики G_2 не забезпечують, наприклад, природного (*що володіє штучним інтелектом*) опису для так званих непроектних конструкцій, тобто для конструкцій з розривними складовими (або з перетином  , 

тощо, або обрамленням  ,  стрілок синтаксичної залежності) [14–16]. Водночас непроектні конструкції є в різних мовах:

Укр. 

[7, 10, 18–20, 32, 33, 36, 45].

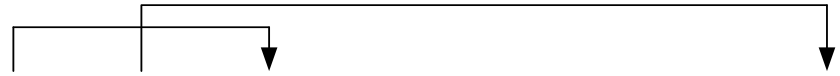
Рос. 

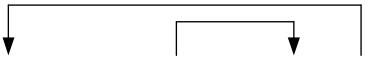
[6, 9, 14–16].

Англ. 

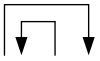
[1, 34, 38–41, 49–57, 60].

Нім.


... die Tatsache, daß die Menschen die Fähigkeit besitzen, Verhältnisse der objektiven Realität in Aussagen wiederzuspiegeln.
[25, 30, 31, 35, 42].

Фр. 

[14–16].

Сербо-хорв. 

[14–16].

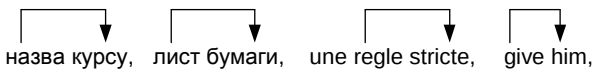
Угор. 

[14–16].

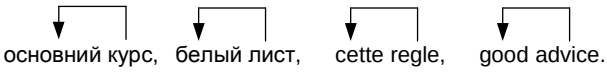
Для опису будови подібних фраз у термінах складових (а граматики G_2 описують синтаксичну структуру саме так) необхідно використовувати розривні складові: всі слова, залежні від одного і того самого слова, повинні утворювати (разом з ним) одну складову, а це за відсутності проективності обов'язково приведе до появи розривних складових (*до цієї поїздки... інтерес*, а theorem ... which describes the properties of this function і т. д.). Проте системи складових граматики, що приписуються фразам граматикою G_2 і, більш того, будь-якою граматикою безпосередніх складових, розривних складових містити не можуть.

Розглянемо два спеціальні випадки граматики G_2, еквівалентні граматикам G_3.

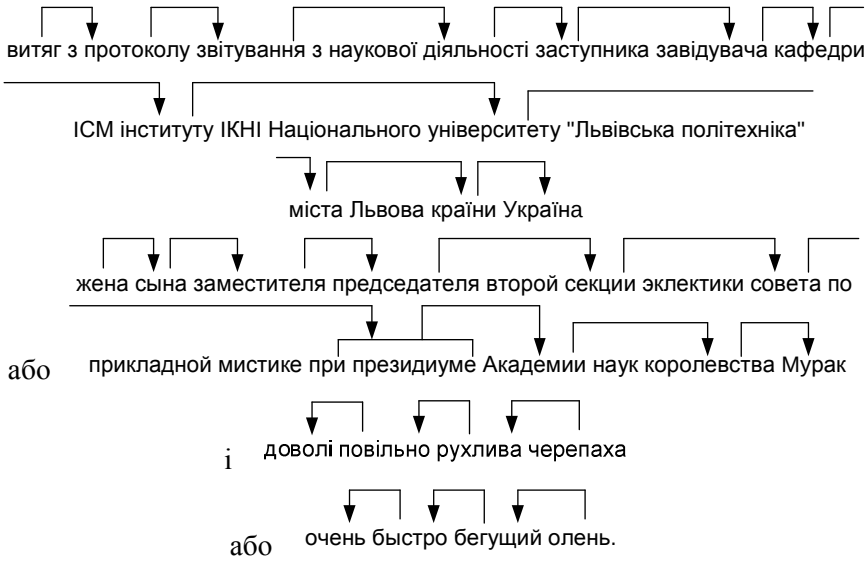
Перший випадок. У природних мовах можливе розміщення залежних слів праворуч від головного (праве підпорядкування) [14–16]:



або ліворуч від головного (ліве підпорядкування) [14–16]:



Як праве, так і ліве підпорядкування може бути послідовним [7, 10, 18–20, 32, 33, 36, 45]:



Залежно від мови та або інша конструкція з послідовним підпорядкуванням праворуч або ліворуч може бути теоретично необмеженою: така, наприклад, конструкція з послідовним підпорядкуванням родових відмінків в українській мові (необмежене праве підпорядкування) і аналогічна конструкція в литовській (де N_p завжди ставиться перед словом, що підпорядковує його, що приводить до необмеженого лівого підпорядкування). Той факт, що мови світу відрізняються і можуть класифікуватися за переважанням в них правого або лівого підпорядкування і, зокрема, залежно від можливості необмеженого послідовного підпорядкування в той або інший бік, був зауважений і досліджений у [63]. Також на цю проблематику – у зв'язку з вживанням граматик G_2 для опису природних мов – звернув увагу В. Інгве [21, 65]). Він зазначив, що існує велика кількість мов (наприклад, англійська, українська, французька тощо), в яких послідовне праве підпорядкування принципово не обмежено, а при лівому підпорядкуванні довжина ланцюжка завжди обмежена через структурні особливості цих мов. Гіпотеза В. Інгве [21] є спробою пояснити це емпіричне спостереження деякими загальними закономірностями людської психіки.

Виявляється, що граматики G_2, що породжує таку мову, має таку цікаву властивість: для будь-якого термінального ланцюжка, що виводиться, наявний такий висновок, у кожному рядку якого всі допоміжні символи збираються в правому кінці, займаючи не більше ніж K останніх місць (K – константа, фіксована для цієї граматики, тобто одна і та сама для всіх виводів в ній). Для того, щоб граматика G_2 мала цю властивість, обмеженості послідовного лівого підпорядкування недостатньо. Необхідне виконання ряду сильніших і складно формульованих вимог [26], із яких випливає, наприклад, обмеженість правого паралельного підпорядкування  і послідовного вкладення типу  [48].

Якщо кожний рядок виведення поділити на дві частини: ліву – одні термінальні символи до першого допоміжного символу X – і праву – від X включно до кінця (у правій частині можуть міститися і термінальні символи), то права частина завжди міститиме не більше ніж K символів. Ліва частина змістовно інтерпретується як вже «виданий» шматок породжуваного ланцюжка (на наступних кроках виведення цей шматок більше не піддається жодній переробці), а права – як робоча ділянка, яку граматика повинна тримати в пам'яті. Отже, число K є не що інше, як максимальний обсяг пам'яті, необхідний для породження будь-якого ланцюжка в цій граматиці (тобто знайдеться ланцюжок, що не породжується за обсягу пам'яті $< K$). Це число збігається з максимальною довжиною ланцюжка послідовних лівих підпорядкувань, можливого у цій мові: так, якщо в якійсь мові не буває більше трьох послідовних підпорядкувань ліворуч, то при породженні цієї мови для будь-якого ланцюжка можна побудувати таке виведення, в якому не виникає необхідності запам'ятовувати більше трьох символів відразу. Такий зв'язок між допустимою глибиною лівого підпорядкування і обсягом пам'яті встановив В. Інгве [21, 65, 66]. Проілюструємо сказане прикладом, а саме розглянемо граматику G_2, що породжує деякі іменні групи української мови [7, 10, 18–20, 32, 33, 36, 45], в яких праве підпорядкування не обмежене, а глибина лівого не перевершує два.

Схема граматики G_2

Приклад 1.

$\tilde{N}_{x,y,z} \to N_{x,y,z}\tilde{N}_{x',y',p}$

$\tilde{N}_{x,y,z} \to \tilde{A}_{x,y,z}\tilde{N}_{x,y,z}$

$\tilde{A}_{x,y,z} \to \{дуже, достатньо, точно, просто, суттєво, ...\}\, A_{x,y,z}$

$\tilde{N}_{x,y,z} \to N_{x,y,z}$

$\tilde{A}_{x,y,z} \to A_{x,y,z}$

$N_{ж,y,z} \to система_{y,z}, ...$

$N_{ч,y,z} \to запит_{y,z}, користувач_{y,z}, ресурс_{y,z}, бізнес_{y,z}, ...$

$A_{x,y,z} \to інформаційний_{x,y,z}, простий_{x,y,z}, ...$

(Тут не враховано особливостей узгодження A з одушевленим $N_{x,y,н}$). Наведемо приклад виведення в граматиці G_2 [7, 10, 18–20, 32, 33, 36, 45]:

$\tilde{N}_{ч,од,н}$

$\tilde{A}_{ч,од,н}\tilde{N}_{ч,од,н}$

досить $A_{ч,од,н}\tilde{N}_{ч,од,н}$

досить простий $A_{ч,од,н}\tilde{N}_{ч,од,н}$

досить простий інформаційний $\tilde{N}_{ч,од,н}\tilde{N}_{ч,од,р}$

доволі простий інформаційний $N_{ч,од,н}\tilde{N}_{ч,од,р}$

доволі простий інформаційний запит $\tilde{N}_{ч,од,р}$

доволі простий інформаційний запит $\tilde{N}_{ч,од,р}\tilde{N}_{ч,од,р}$

доволі простий інформаційний запит $N_{ч,од,р}\tilde{N}_{ч,од,р}$

доволі простий інформаційний запит користувача $\tilde{N}_{ч,од,р}$

доволі простий інформаційний запит користувача $\tilde{N}_{ч,од,р}\tilde{N}_{ж,од,р}$

доволі простий інформаційний запит користувача $N_{ч,од,р}\tilde{N}_{ж,од,р}$

доволі простий інформаційний запит користувача ресурсу $\tilde{N}_{ж,од,р}$

доволі простий інформаційний запит користувача ресурсу $\tilde{N}_{ж,од,р}\tilde{N}_{ч,од,р}$

доволі простий інформаційний запит користувача ресурсу $N_{ж,од,р}\tilde{N}_{ч,од,р}$

доволі простий інформаційний запит користувача ресурсу системи $\tilde{N}_{ч,од,р}$

доволі простий інформаційний запит користувача ресурсу системи $N_{ч,од,р}$

доволі простий інформаційний запит користувача ресурсу системи бізнесу

Приклад 2.

$\tilde{N}_{x,y,z} \to N_{x,y,z}\tilde{N}_{x',y',p}$

$\tilde{N}_{x,y,z} \to \tilde{A}_{x,y,z}\tilde{N}_{x,y,z}$

$\tilde{A}_{x,y,z} \to \{дуже, доволі, точно, просто, суттєво, ...\}A_{x,y,z}$

$\tilde{N}_{x,y,z} \to N_{x,y,z}$

$\tilde{A}_{x,y,z} \to A_{x,y,z}$

$N_{ж,y,z} \to школа_{y,z}, ...$

$N_{ч,y,z} \to сміх_{y,z}, учень_{y,z}, Львів_{y,z}, ...$

$N_{с,y,z} \to місто_{y,z}, ...$

$A_{x,y,z} \to веселий_{x,y,z}, дитячий_{x,y,z}, ...$

(Тут не враховано особливостей узгодження A з одушевленим $N_{x,y,н}$.) Наведемо приклад виведення в граматиці G_2 [7, 10, 18–20, 32, 33, 36, 45]:

$\tilde{N}_{ч,од,н}$

$\tilde{A}_{ч,од,н}\tilde{N}_{ч,од,н}$

дуже $A_{ч,од,н}\tilde{N}_{ч,од,н}$

дуже веселий $A_{ч,од,н}\tilde{N}_{ч,од,н}$

дуже веселий дитячий $\tilde{N}_{ч,од,н}\tilde{N}_{ч,од,р}$

дуже веселий дитячий $N_{ч,од,н}\tilde{N}_{ч,од,р}$

дуже веселий дитячий сміх $\tilde{N}_{ч,од,р}$

дуже веселий дитячий сміх $\tilde{N}_{ч,од,р}\tilde{N}_{ж,од,р}$

дуже веселий дитячий сміх $N_{ч,од,р}\tilde{N}_{ж,од,р}$

дуже веселий дитячий сміх учня $\tilde{N}_{ж,од,р}$

дуже веселий дитячий сміх учня $\tilde{N}_{ж,од,р}\tilde{N}_{с,од,р}$

дуже веселий дитячий сміх учня $N_{ж,од,р}\tilde{N}_{с,од,р}$

дуже веселий дитячий сміх учня школи $\tilde{N}_{с,од,р}$

дуже веселий дитячий сміх учня школи $\tilde{N}_{с,од,р}\tilde{N}_{ч,од,р}$

дуже веселий дитячий сміх учня школи $N_{с,од,р}\tilde{N}_{ч,од,р}$

дуже веселий дитячий сміх учня школи міста $\tilde{N}_{ч,од,р}$

дуже веселий дитячий сміх учня школи міста $N_{ч,од,р}$

дуже веселий дитячий сміх учня школи міста Львова

У цьому виведенні обсяг пам'яті дорівнює двом: жоден проміжний ланцюжок не містить понад два допоміжні символи. Той самий ланцюжок можна було б породити і по-іншому, використовуючи більший обсяг пам'яті, наприклад, спочатку отримати з $\tilde{N}_{ч,од,н}$ ланцюжок

$$\textit{дуже}\ A_{ч,од,н}\ A_{ч,од,н}\ N_{ч,од,н}\ N_{ч,од,р}\ N_{ж,од,р}\ N_{с,од,р}\ N_{ч,од,р}$$

а вже з неї наш термінальний ланцюжок. Для нас проте важливий *необхідний* обсяг пам'яті, тобто такий, що з меншим обсягом отримати цей ланцюжок неможливо. Саме цей обсяг і дорівнює тут двом.

Можна довести, що і будь-який термінальний ланцюжок, що виводиться в G_2, може бути породжений з об'ємом пам'яті ≤ 2. Доказ оснований на дуже простому міркуванні: «хороше» виведення треба проводити так, щоб для кожного іменника спочатку видавалися в термінальному вигляді його ліві залежні, і тільки потім іменна група праворуч розгорталася.

Теорема 1. Граматика G_2 описаного типу (із обмеженою пам'яттю) завжди еквівалентна деякій граматиці G_3 [14–16]. Це неважко довести (оскільки права частина рядка виведення, що складається з K символів, кодується одним новим допоміжним символом). Отже, в разі мов із обмеженою глибиною лівого підпорядкування граматика G_2 з обмеженою пам'яттю, еквівалентні граматикам G_3 і близькі до них за побудовою виведень, тобто влаштовані набагато простіше, ніж довільні граматики G_2, виявляються не лише принципово достатніми, але і вельми зручними – вони забезпечують достатньо природний опис.

Другий випадок. Бувають проте мови, в яких необмежену глибину має не лише праве, але і ліве послідовне підпорядкування. Такою мовою є, наприклад, угорська, де ліве підпорядкування не обмежене завдяки препозитивним поширеним визначенням, а праве – завдяки, наприклад, підрядним реченням з *який* (Будинок, який побудував Джек) [14-16]. Див. приклад з новели Г. Фехера – жартівливий тост, наведений в роботі [68].

1. Kivánom, hogy valamint az agyag[23] цlelx karjai[22] közül kibontakozni[21] akary[20] kocsikerйk[19] rettentx nyikorgбsбtyl[18] megriadt[17] juhбszkutya[16] bundбjбba[15] kapaszkody[14] kullancs[13] kidülledt fйlszemйbxl[12] alбcseppent[11] kцnnyeseppben[10] visszatьkrцzxdx[9] holdvilág fйnyйtxl[8] illuminбlt[7] rablylovagvбr[6] felvonyhidjбbyl[5] kiбlly[4] vasszegek[3] kohйziys erejйnek[2] hatбsa[1] évszázadokra összetartja annak materiáját, aképpen tartsa össze ezt a társaságot az igaz szeretet.

2. Я хочу, аби справжнє кохання скріпило цю компанію так, як на століття скріплює матеріал мосту дія[1] єднальної сили[2] цвяхів[3], що торчать[4] з підіймального мосту[5] розбійницького феодального замку[6], освітленого[7] місячним світлом[8], що відображається[9] в краплині[10], яка витікає[11] з ока[12] кліща[13], що вчепився[14] в шерсть[15] вівчарки[16], наполоханої[17] жахливим скрипом[18] колес возу[19], що прагнуть[20] вирватися[21] з обіймів[22] грязюки[23] [7, 10, 18–20, 32, 33, 36, 45].

Ця фраза з художнього тексту має глибину 22 і є абсолютно правильною з граматичного погляду (точно такою мірою, як і її український переклад). Більш того, ніщо не заважає продовжити ланцюг визначень ліворуч *ad libitum*.

Для породження мов із такою властивістю можна запропонувати ще один особливий тип граматики G_2, у деякому аспекті загальнішої, ніж граматики G_2 із обмеженою пам'яттю, розглянуті вище. Перш за все сформулюємо точніше, які мови маємо на увазі. Це мови, в яких можливо необмежена кількість послідовне підлеглих зліва направо конструкцій $X_1X_2...X_i...$ (необмежене праве підпорядкування), і при цьому в кожній з конструкцій X_i можливо необмежене ліве підпорядкування – послідовність конструкцій $...X_{ij}...X_{i3}X_{i2}X_{i1}$; проте усередині конструкцій X_{ij} подальше необмежене розгортання неможливе. Стосовно угорської мови X_i можна розуміти як прості речення, що є кожне (окрім першого) додатковим визначальним до попереднього, а X_{ij} – як препозитивні дієприкметникові звороти.

Розглянемо граматику $\Gamma' = \langle V', V_1', I', S' \rangle$, основний словник якої V' складається з n символів $A_1, A_2, ..., A_n$ і правила якої мають вигляд $X \to YA_i$ або $X \to A_i$, де X і Y належать до V_1' [14–16]. Нехай кожному з символів A_i відповідає деяка регулярна граматика $\Gamma_i' = \langle V, V_1^i, A_i, S_i \rangle$, де V – основний словник, загальний для всіх Γ_i', V_1^i – допоміжний словник, що не містить жодних символів із V' і V_1', крім A_i; A_i – початковий символ; правила схеми S_i мають вигляд $C \to dD$ або $C \to c$ (тут, як і в інших прикладах, заголовними латинськими літерами позначено допоміжні символи, а рядковими – основні). При цьому вважатимемо, що допоміжні словники граматик Γ_i' попарно не перетинаються.

Граматика Γ' дуже близька до автоматної, відрізняючись від неї лише напрямом розгортання (під напрямом розгортання тут розуміємо напрям, в якому «викидаються» термінальні символи, наприклад, $C \to dD$ – ліве розгортання) породжуваного ланцюжка; по суті вона є автоматною з точністю до дзеркальної симетрії. Отже, маємо справу з однією квазірегулярною граматикою правого розгортання і з n регулярними граматиками лівого розгортання.

Розглянемо тепер об'єднання всіх цих граматик, точніше, граматику Γ, в якої основний словник – V (той самий, що у всіх Γ_i'), допоміжний словник – $V_1 = V' \cup V_1' \cup V_1^1 \cup V_1^2 \cup ... \cup V_1^n$ (тобто об'єднання допоміжних словників всіх граматик $\Gamma' \Gamma_1' \Gamma_2', ..., \Gamma_n'$ і основний словник граматики Γ'), початковий символ — I (той самий, що у Γ'), а схема є поєднанням схем всіх граматик Γ', Γ_1', $\Gamma_2', ..., \Gamma_n'$. Ця граматика Γ є спеціальною контекстно-вільною граматикою, яку можна назвати *контекстно-вільною граматикою з незалежним двобічним розгортанням*. Те, що ця граматика не є автоматною, очевидно хоча би тому, що деякі її правила (правила схеми S) мають в правих частинах по два допоміжні символи. Основні символи граматики Γ' (т. е. $A_1, A_2, ..., A_n$) в граматиці Γ є допоміжними, так що правила вигляду $X \to YA_i$ в межах Γ є не автоматними. Але граматика Γ еквівалентна автоматній. Наведемо приклад (схеми) такої граматики.

$$S' = \begin{cases} I \to BA_1 \\ B \to CA_1 \\ C \to BA_2 \\ C \to DA_3 \\ D \to DA_4 \\ D \to A_2 \end{cases} \qquad S_1 = \begin{cases} A_1 \to bP_1 \\ P_1 \to aQ_1 \\ Q_1 \to aQ_1 \\ Q_1 \to c \end{cases} \qquad S_2 = \{A_2 \to d \quad S_3 = \begin{cases} A_3 \to aP_3 \\ A_3 \to bQ_3 \\ A_3 \to cR_3 \\ P_3 \to a \\ Q_3 \to b \\ R_3 \to dR_3 \\ R_3 \to eR_3 \\ R_3 \to d \end{cases} \qquad S_4 = \begin{cases} A_4 \to cP_4 \\ P_4 \to b \end{cases}$$

Граматика введеного Гладким типу працює так. Спочатку породжуваний ланцюжок необмежено розгортається зліва направо завдяки символам A_i (які можуть інтерпретуватися, наприклад, як синтаксичні групи або речення); це робиться правилами S'. Потім будь-яке з A_i може (правилами S_i) необмежено розгортатися справа наліво – в ланцюжок термінальних символів (які можна інтерпретувати як слова). Такий процес породження зручний в таких, наприклад, випадках, як угорські фрази розглянутого вище типу.

Теорема 2. Кожна контекстно-вільна граматика з незалежним двостороннім розгортанням еквівалентна деякій регулярній граматиці [14–16].

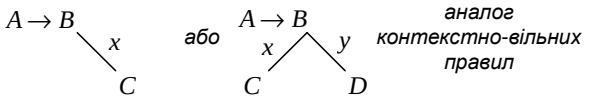
Необмежені граматики типу 0 є лише окремим випадком загального поняття граматики. Проте вони безумовно достатні для опису будь-яких природних мов в повному обсязі. Будь-яка природна мова (множина правильних фраз) є легко розпізнаваною множиною. Це означає існування доволі простого алгоритму розпізнавання правильності фраз. Якщо ж мова розпізнається алгоритмом з вказаним обмеженням на обсяг пам'яті, то вона може бути породженою граматикою, де для будь-якого термінального ланцюжка довжини n, що виводиться, існує таке виведення, в якому жоден проміжний ланцюжок не перевищує за довжиною числа Kn (K – деяка константа). Така граматика є *граматикою з обмеженим розтягуванням*, де ємнісна сигнальна функція не більша за лінійну. Для будь-якої граматики з обмеженим розтягуванням можна побудувати еквівалентну їй граматику G_0, яка здатна описувати множину правильних фраз будь-якої природної мови, тобто породжувати будь-які правильні фрази цієї мови, не породжуючи при цьому жодних неправильних. Обидві конструкції, наведені як приклади непридатності контекстно-вільних граматик, легко описуються граматикою G_0.

Недоліки *методу виведення* граматики G_0 зводяться до трьох пунктів.

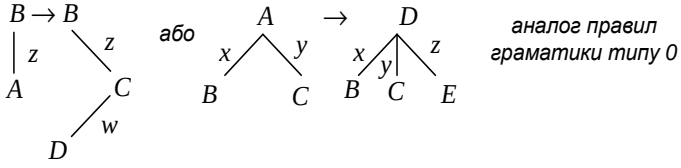
1. За їх допомогою не можливо природно описати фрази з розривними складовими.

2. Граматика G_0 містить лише правила утворення мовних виразів, наприклад, словоформ або фраз. Граматика задає правильні вирази на відміну від неправильних.

3. Граматики G_0 будують речення відразу з певною послідовністю слів, яку ці речення повинні мати в остаточному вигляді. При цьому породжуваному реченню зіставляється синтаксична структура у формі впорядкованого дерева, тобто дерева, де між вузлами, окрім відношення підпорядкування, що задається самим деревом, є ще і відношення лінійного порядку (правіше – лівіше). Отже, в синтаксичній структурі граматики G_0 не розчленовано два абсолютно різних за природою, хоча і зв'язаних між собою відношення: синтаксичне підпорядкування і лінійне взаєморозташування. Але охарактеризувати синтаксичну структуру – це вказати відношення синтаксичного підпорядкування. Що ж до відношення лінійного порядку, то воно характеризує не структуру, а саму фразу. Послідовність слів залежить від синтаксичної структури і визначається обов'язково з її обліком і є відносно до неї чимось похідним, вторинним. Доцільно видозмінити поняття граматики, що породжує, так, щоб ліві і праві частини правил підстановки були не лінійно впорядкованими ланцюжками, а наприклад, деревами (без лінійної впорядкованості), що характеризують синтаксичні відношення [14–16]. Тоді правила набувають вигляду:



або



Риски з індексами зображають синтаксичні зв'язки різних типів; літери $A, B, C, \ldots$ – синтаксичні категорії. NB: взаємне розташування символів одного рівня підпорядкування не має жодного значення і є на даній схемі випадковим 

; означає те ж саме, що і 

[14–16].

У результаті отримують обчислення синтаксичних структур (а не фраз) мови. Це обчислення є частиною граматики, що породжує. Іншу частину цієї граматики становить обчислення, яке для будь-якої даної синтаксичної структури задає (з врахуванням яких-небудь інших факторів, наприклад, в українській мові – з обов'язковим обліком логічного виділення тощо) всі можливі для неї лінійні послідовності слів. Тоді знімається проблема розривних складових. Із виведення речення в регулярній граматиці неможливо отримати природне подання структури безпосередніх складових цього речення. Тобто, регулярні граматики дають деяку структуру складових, як і взагалі всі граматики безпосередніх складових, однак, ці складові зазвичай мають формальний характер.

Текстовий контент C_2 (стаття, коментар, книга тощо) містить значний обсяг даних природною мовою, частина яких є абстрактною. Текст подають як об'єднану за змістом послідовність знакових одиниць, основними властивостями якої є інформаційна, структурна та комунікативна зв'язність/цілісність, що відображає змістовну/структурну сутність тексту. Методом опрацювання тексту є лінгвістичний аналіз змісту (наприклад, коментарі, форуми, статті тощо). Процес опрацювання тексту поділяє контент на лексеми за допомогою кінцевих автоматів. Як функціонально-семантико-структурна єдність текст має правила побудови, виявляє закономірності змістовного та формального з'єднання складових. Зв'язність проявляється через зовнішні структурні показники та формальну залежність компонентів тексту, а цілісність – через тематичну, концептуальну та модальну залежність. Цілісність веде до змістовної та комунікативної організації тексту, а зв'язність – до форми, структурної організації. Оператор виявлення ключових слів комерційного контенту $\alpha_4 : (C_2, U_K, T) \to C_3$ є відображенням комерційного контенту C_2 в новий стан, який відрізняється від попереднього стану наявністю множини ключових слів, що загально описують його зміст. Під час аналізу досліджують багаторівневу структуру контенту: лінійну послідовність символів; лінійну послідовність морфологічних структур; лінійну послідовність речень; мережу взаємопов'язаних єдностей (алг. 1). Під час аналізу досліджують багаторівневу структуру текстового контенту: лінійну послідовність символів; лінійну послідовність морфологічних структур; лінійну послідовність речень; мережу взаємопов'язаних єдностей (алг. 1).

Алгоритм 1. Лінгвістичний аналіз текстового комерційного контенту

Етап 1. Граматичний аналіз текстового контенту C_2.

Крок 1. Поділ текстового комерційного контенту C_2 на речення та абзаци.

Крок 2. Поділ ланцюжка символів контенту C_2 на слова.

Крок 3. Виділення цифр, чисел, дат, незмінних оборотів і скорочень контенту C_2.

Крок 4. Видалення нетекстових символів контенту C_2.

Крок 5. Формування та аналіз лінійної послідовності слів із службовими знаками для контенту C_2 (алг. 3).

Етап 2. Морфологічний аналіз текстового контенту C_2.

Крок 1. Отримання основ (словоформ із відрубаними закінченнями).

Крок 2. Для кожної словоформи формується граматична категорія (колекція граматичних значень: рід, відмінок, відмінювання тощо).

Крок 3. Формування лінійної послідовності морфологічних структур.

Етап 3. Синтаксичний аналіз $\alpha_4 : (C_2, U_K, T) \to C_3$ текстового контенту C_2 (алг. 2).

Етап 4. Семантичний аналіз текстового контенту C_3.

Крок 1. Слова співвідносяться з семантичними класами із словника.

Крок 2. Відбір потрібних для певного речення морфосемантичних альтернатив.

Крок 3. Зв'язування слів у єдину структуру.

Крок 4. Формування упорядкованої множини записів суперпозицій з базисних лексичних функцій і семантичних класів. Точність результату визначається повнотою/коректністю словника.

Етап 5. Референційний аналіз для формування міжфразових єдностей.

Крок 1. Контекстний аналіз текстового комерційного контенту C_3. За його допомогою реалізується дозвіл локальних референцій (цей, який, його) і виділення висловлювання – ядра єдності.

Крок 2. Тематичний аналіз. Поділ висловлювань на тему і рему виділяє тематичні структури, які використовують, наприклад, для формування дайджесту.

Крок 3. Визначають регулярну повторюваність, синонімізацію та повторну номінацію ключових слів; тотожність референції, тобто співвідношення слів з предметом зображення; наявність імплікації, основаної на ситуативних зв'язках.

Етап 6. Структурний аналіз текстового контенту C_3. Передумовами використання є високий ступінь збігу термінів єдності, дискурсивна одиниця, речення семантичною мовою, висловлювання і елементарна дискурсивна одиниця.

Крок 1. Виявлення базового набору риторичних зв'язків між єдностями контенту.

Крок 2. Побудова нелінійної мережі єдностей. Відкритість набору зв'язків припускає його розширення та адаптацію для аналізу структури текстів C_3.

Синтаксичні аналізатори працюють в два етапи: ідентифікують змістовні лексеми та створюють дерево розбору (алг. 2). Текст реалізує структурно подану діяльність, що передбачає суб'єкт і об'єкт, процес, мету, засоби і результат, які відображаються в змістовно-структурних, функціональних, комунікативних показниках. Одиницями внутрішньої організації структури тексту є алфавіт, лексика (парадигматика), граматика (синтагматика), парадигми, парадигматичні відношення, синтагматичні відношення, правила ідентифікації, висловлювання, міжфразова єдність та фрагменти-блоки. На композиційному рівні виділяють речення, абзаци, параграфи, розділи, глави, підглави, сторінки тощо, які, крім речення, опосередковано пов'язані з внутрішньою структурою, тому не розглядаються. За допомогою бази даних (бази термінів/морфем і службових частин мови) та визначених правил аналізу тексту шукають термін. Синтаксичні аналізатори працюють в два етапи: ідентифікують змістовні лексеми та створюють дерево розбору (алг. 2).

Алгоритм 2. Синтаксичний аналізатор комерційного контенту

Етап 1. Ідентифікація змістовних лексем $U_{K1} \in U_K$ для комерційного контенту C_2.

Крок 1. Визначення ланцюжка термів у вигляді речення.

Крок 2. Ідентифікація іменної групи за допомогою словника основ.

Крок 3. Ідентифікація дієслівної групи за допомогою словника основ.

Етап 2. Створення дерева розбору зліва направо. Виведення дерева полягає в розгортанні одного з символів попереднього ланцюжка послідовності лінгвістичних змінних або в заміні його іншим, інші ж символи переписуються без зміни. При розгортанні, замінювані/переписувані символи (*предки*) з'єднують безпосередньо з символами, які отримують в результаті розгортання, заміни або переписування (*нащадками*), та отримують дерево складових, або синтаксичну структуру для змісту комерційного контенту.

Крок 1. Розгортання іменної групи. Розгортання дієслівної групи.

Крок 2. Реалізація синтаксичних категорій словоформами.

Етап 3. Визначення множини ключових слів $\alpha_4 : (C_2, U_K, T) \to C_3$ для контенту C_2.

Крок 1. Визначення термів $Noun \in U_{K1}$ – іменників, словосполучень іменників або прикметника з іменником серед множини слів текстового контенту.

Крок 2. Розрахунок унікальності $Unicity$ для термів $Noun \in U_{K1}$.

Крок 3. Розрахунок $NumbSymb \in U_{K3}$ (кількість знаків без пробілів) для $Noun \in U_{K1}$ при $Unicity \geq 80$.

Крок 4. Розрахунок $UseFrequency \in U_{K2}$ – частоти появи ключових слів комерційного контенту. Для термів з $NumbSymb \leq 2000$ частота $UseFrequency$ є в межах $(6;8]$%, з $NumbSymb \geq 3000$ – $[2;4)$%, з $2000 > NumbSymb < 3000$ – $[4;6]$%.

Крок 5. Розрахунок $BUseFrequency$ – частота появи ключових слів на початку тексту, $IUseFrequency$ – частота появи ключових слів в середині тексту, $EUseFrequency$ – частота появи ключових слів в кінці тексту комерційного контенту.

Крок 6. Порівняння значень $BUseFrequency$, $IUseFrequency$ та $EUseFrequency$ для розстановки пріоритетів. Ключові слова з більшими значеннями $BUseFrequency$ мають більший пріоритет, ніж ключові слова з більшим значенням $EUseFrequency$.

Крок 7. Сортування ключових слів за пріоритетом.

Етап 4. Заповнення бази пошукових образів контенту C_3, тобто атрибутів $KeyWords \in U_{K4}$ – ключові слова, $Unicity$ – унікальність ключових слів ≥ 80, $Noun$ – терм, $NumbSymb$ – кількість знаків без пробілів, $UseFrequency$ – частота вживання ключових слів, $BUseFrequency$ – частота вживання ключових слів на початку тексту, $IUseFrequency$ – частота вживання ключових слів в середині тексту, $EUseFrequency$ – частота вживання ключових слів в кінці тексту.

Ключові слова тематики комерційного контенту C_2 з фрагмента тексту виявляють за допомогою процесів, поданих на рис. 1. Текст реалізує структурно подану діяльність, що передбачає суб'єкт і об'єкт, процес, мету, засоби і результат, які відображаються в змістовно-структурних, функціональних, комунікативних показниках. Одиницями внутрішньої організації структури тексту є алфавіт, лексика (парадигматика), граматика (синтагматика), парадигми, парадигматичні відношення, синтагматичні відношення, правила ідентифікації, висловлювання, міжфразова єдність та фрагменти-блоки. На композиційному рівні виділяють речення, абзаци, параграфи, розділи, глави, підглави, сторінки тощо, які, крім речення, опосередковано пов'язані з внутрішньою структурою, тому не розглядаються. За допомогою бази даних (бази термінів/морфем і службових частин мови) та визначених правил аналізу тексту шукають термін.

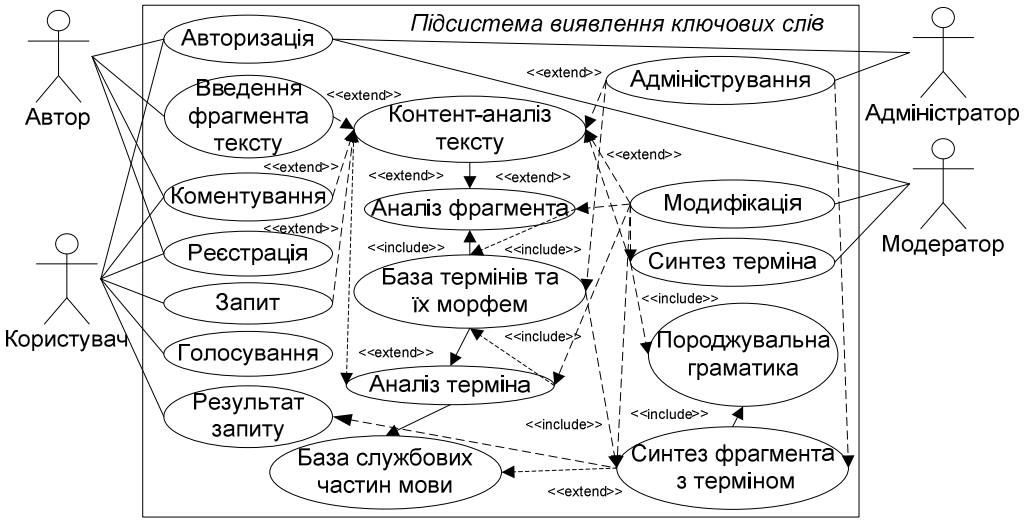


Рис. 1. Діаграма варіантів виявлення ключових слів

За правилами породжувальної граматики корегують термін згідно із правилами його вживання у контексті. Речення задають межі дії знаків пунктуації, анафоричних і катафоричних посилань. Семантика тексту зумовлена комунікативним завданням передавання даних. Структура тексту визначається внутрішньою організацією одиниць тексту і закономірностями їх взаємозв'язку. За допомогою синтаксичного аналізу текст оформляють у структуру даних, наприклад, в дерево, яке відповідає синтаксичній структурі вхідної послідовності і якнайкраще підходить для подальшого опрацювання. Після аналізу фрагмента тексту і терміна синтезують новий термін як ключове слово тематики контенту, використовуючи базу термінів та їх морфем. Потім синтезують терміни для формування нового ключового слова, використовуючи базу службових частин мови. Принцип виявлення ключових слів за термами ґрунтується на законі Зіпфа і зводиться до вибору слів із середньою частотою появи (найвживаніші слова ігнорують через стоп-словники, а рідкісні слова не враховують).

За змістовний аналіз контенту відповідає процес витягування граматичних даних зі слова за допомогою графемного аналізу та корегування результатів морфологічного аналізу за допомогою аналізу граматичного контексту лінгвістичних одиниць (алг. 3).

Алгоритм 3. Рубрикація текстового комерційного контенту

Етап 1. Поділ комерційного контенту C_3 на блоки.

Крок 1. Подання на вхід блоку побудови дерева блоків комерційного контенту C_3.

Крок 2. Створення нового блоку в таблиці блоків.

Крок 3. Накопичення символів до символу нового рядка.

Крок 4. Перевірка на наявність крапки перед символом нового рядка. Якщо є, то перехід до кроку 5, якщо ні, то збереження послідовності у таблиці, розбір нового блоку контенту C_3 та перехід на крок 3.

Крок 5. Перевірка наявності кінця тексту для контенту C_3. Якщо кінець тексту, то перехід до кроку 6, якщо ні, то зберігається накопичена послідовність у таблицю, розбір нового блоку контенту C_3 та перехід до кроку 2.

Крок 6. Отримання на виході дерева блоків контенту C_3 у вигляді таблиці $U_{CT}^{B} \in U_{CT}$.

Етап 2. Поділ блоку на речення зі збереженням структури контенту C_3.

Крок 1. На вхід подається таблиця блоків $U_{CT}^{B} \in U_{CT}$. Створення таблиці речень $U_{CT}^{R} \in U_{CT}$ із зв'язком за полем Код_розділу типу *n-to-*1 із таблицею блоків контенту C_3.

Крок 2. Створення нового речення в таблиці речень $U_{CT}^{R} \in U_{CT}$.

Крок 3. Накопичення символів до крапки, крапки з комою або символу нового рядка.

Крок 4. Перевірка на наявність скорочення. Якщо скорочення, то перехід до кроку 5, якщо ні, то збереження послідовності у таблиці $U_{CT}^{R} \in U_{CT}$, розбір нового речення та перехід до кроку 2.

Крок 5. Перевірка наявності кінця тексту блоку для контенту C_3. Якщо кінець тексту, то перехід до кроку 6, якщо ні, то збереження послідовності у таблиці $U_{CT}^{R} \in U_{CT}$, розбір нового речення та перехід до кроку 2.

Крок 6. Отримують на виході дерево речень у вигляді таблиці $U_{CT}^{R} \in U_{CT}$.

Крок 7. Перевірка наявності кінця тексту для контенту C_3. Якщо кінець тексту, то перехід до кроку 8, якщо ні, то розбір нового блоку та перехід до кроку 1.

Крок 8. Отримання на виході дерева речень у вигляді таблиць $U_{CT}^{R} \in U_{CT}$.

Етап 3. Поділ речень на лексеми із вказанням належності до речень $U_{CT}^{L} \in U_{CT}$.

Крок 1. Формування на основі таблиці речень таблиці лексем $U_{CT}^{L} \in U_{CT}$ із полями Код_лексеми (унікальний ідентифікатор), Код_речення (число, що дорівнює коду речення із лексемою), Номер_лексеми (число, яке дорівнює номеру лексеми в реченні), Текст (текст лексеми).

Крок 2. Подання на вхід для розбору на лексеми речення з таблиці речень $U_{CT}^{R} \in U_{CT}$.

Крок 3. Створення нової лексеми в таблиці лексем $U_{CT}^{L} \in U_{CT}$.

Крок 4. Накопичення символів до крапки, пропусків або кінця речення та збереження в таблиці лексем.

Крок 5. Перевірка кінця речення. Якщо так, то перехід до кроку 6, якщо ні, то збереження накопиченої послідовності у таблицю $U_{CT}^{L} \in U_{CT}$, розбір нової лексеми та перехід до кроку 3.

Крок 6. Проведення синтаксичного аналізу на основі вихідних даних (алг. 2).

Крок 7. Проведення морфологічного аналізу на основі даних, одержаних на виході.

Етап 4. Визначення тематики комерційного контенту $U_{CT}^{T} \in U_{CT}$.

Крок 1. Побудова ієрархічної структури властивостей $U_{CT}^{T} \in U_{CT}$ кожної лексичної одиниці тексту, що містить граматичну та семантичну інформацію.

Крок 2. Формування лексикону з ієрархічною організацією типів властивостей, де кожен тип-нащадок успадковує і перевизначає властивості предка.

Крок 3. Уніфікація – базовий механізм побудови синтаксичної структури.

Крок 4. Визначення ключових слів *KeyWords* комерційного контенту $C_4 = \alpha_5(\alpha_4(C_2, U_K), U_{CT})$ при $U_{CT} = \{U_{CT1}, U_{CT2}, U_{CT3}, U_{CT4}\}$, де U_{CT} – колекція умов рубрикації, U_{CT1} – множина тематичних ключових слів зі словника, U_{CT2} – множина частот вживання ключових слів у комерційному контенті, U_{CT3} – множина залежностей вживання ключових слів різних тематик (коефіцієнти визначає модератор згідно із належністю ключового слова до певної тематики в межах $[0,1]$), U_{CT4} – множина частот вживання тематичних ключових слів у контенті (алг.2).

Крок 5. Визначення $U_{Ct}^{T} \in U_{Ct}$ з *TKeyWords* – тематичні ключові слова в множині *KeyWords* для . *Topic* – тема контенту та *Category* – категорія контенту.

Крок 6. Визначення *FKeyWords* – частота вживання ключових слів та *QuantitativeryTKey* – частота вживання тематичних ключових слів у тексті комерційного контенту.

Крок 7. Визначення *Comparison* – порівняння появи ключових слів різних тематик Розрахунок *CofKeyWords* – коефіцієнт тематичних ключових слів контенту, *Static* – коефіцієнт статистичної важливості термів, *Addterm* – коефіцієнт наявності додаткових термів. Порівняння множини ключових слів контенту з ключовими поняттями тем. Якщо є збіг, то перехід до кроку 9, якщо ні, то перехід до кроку 8.

Крок 8. Формування нової рубрики з набором ключових понять аналізованого контенту C_4.

Крок 9. Присвоєння визначеній рубриці аналізованого комерційного контенту C_4.

Крок 10. Розрахунок *Location* – коефіцієнт розташування контенту C_4 у тематичній рубриці.

Етап 4. Заповнення бази пошукових образів для атрибутів *Topic* – тема контенту, *Category* – категорія контенту, *Location* – коефіцієнт розташування контенту в тематичній рубриці, *CofKeyWords* – коефіцієнт тематичних ключових слів у контенті, *Static* – коефіцієнт статистичної важливості термів, *Addterm* – коефіцієнт наявності додаткових термів, *TKeyWords* – тематичні ключові слова, *FKeyWords* – частота вживання ключових слів, *Comparison* – порівняння появи ключових слів різних тематик, *QuantitativeryTKey* – частота вживання тематичних ключових слів у тексті комерційного контенту C_4.

Побудова тексту контенту C_4 визначається темою, поданою інформацією, умовами спілкування, завданням повідомлення та стилем викладення. Із семантичною, граматичною та композиційною структурою контенту C_4 пов'язані його стильові/стилістичні характеристики, залежні від індивідуальності автора та підпорядковані тематичній/стильовій домінанті тексту. Процес рубрикації контенту C_4 у вигляді діаграми варіантів подано на рис. 2.

Рис. 2. Діаграма варіантів використання для процесу рубрикації контенту в СЕКК

Основні етапи визначення морфологічних ознак U_{CT} одиниць тексту C_4: визначення граматичних класів слів – частин мови і принципів їх класифікаційного виділення; виокремлення частини семантики слова як морфологічної; обґрунтування набору морфологічних категорій та їх природи; опис сукупності формальних засобів, закріплених за частинами мови та їх морфологічними категоріями. Процес рубрикації $C_4 = \alpha_5(\alpha_4(C_2, U_K), U_{CT})$ через автоматичне індексування складових комерційного контенту C_3 розбита на послідовні блоки: морфологічний аналіз, синтаксичний аналіз, семантико-синтаксичний аналіз лінгвістичних конструкцій та варіювання змістовного запису текстового контенту.

Використано такі способи виразу граматичного значення: синтетичний, аналітичний, аналітико-синтетичний та суплетивний. Граматичні значення узагальнено через однотипні характеристики та підлягають поділу на часткові значення. Для позначення класів однотипних граматичних значень використано поняття граматичної категорії. До морфологічних значень належать категорії роду, числа, відмінка, особи, часу, способу, стану, виду, об'єднані у парадигми для класифікації частин тексту. Об'єктом морфологічного аналізу є структура слова, форми словозміни, способи виразу граматичних значень. Морфологічні ознаки одиниць тексту – це інструменти дослідження зв'язку між лексикою, граматикою, використанням їх у мовленні, парадигматикою (відмінкові форми відмінюваних слів) і синтагматикою (лінійні зв'язки слів, сполучення). Реалізація автоматичного кодування слів тексту, тобто приписування їм кодів граматичних класів, пов'язана з граматичною класифікацією. Морфологічний аналіз містить такі етапи: виділення основи у словоформі; пошук основи у словнику основ; порівняння структури словоформи з даними у словниках основ, коренів, префіксів, суфіксів, флексій. У процесі аналізу ідентифікують значення слів та синтагматичних відношень між словами контенту. Інструментами аналізу є словники основ/флексій/омонімів та статистичних/ синтаксичних словосполучень, зняття лексичної омонімії, семантичний аналіз іменних безприйменникових конструкцій, таблиці семантико-синтаксичного сполучення іменників/прикметників та компонентів прийменникових конструкцій, алгоритми аналізу для визначення послідовностей перевірок і звертань до словника і таблиць; система поділу слів тексту на флексію й основу; тезаурус еквівалентностей для заміни еквівалентних слів одним/кількома номерами понять, які слугують ідентифікаторами змісту замість основ слів; тезаурус у вигляді ієрархії понять для пошуку щодо цього поняття загального/асоційованого з ним поняття; система обслуговування словників. Процес індексування залежить від дескрипторного словника або інформаційно-пошукового тезауруса. Дескрипторний словник має структуру таблиці з трьома колонками: основи слів; набори дескрипторів, приписані кожній основі; граматичні ознаки дескрипторів. Індексування передбачає виділення інформативних словосполучень з тексту; розшифрування абревіатури; заміну слів з основами-дескрипторами на код дескриптора; зняття омонімії.

Висновки і перспективи подальших наукових розвідок

Дослідження застосування математичних методів для аналізу та синтезу текстової інформації природною мовою необхідні для розроблення математичних алгоритмів та комп'ютерних програм опрацювання текстового контенту. Апарат породжувальних граматик, запропонований Н. Хомські, моделює процеси на синтаксичному рівні мови. Виділені структурні елементи речення описують синтаксичні конструкції текстового контенту незалежно від їх змісту. У статті показано особливості процесу синтезу речень різних мов із застосуванням породжувальних граматик. У роботі розглянуто вплив норм та правил мови на процес побудови граматик. Застосування породжувальних граматик має широкі можливості у розробленні та створенні автоматизованих систем опрацювання текстового контенту, для лінгвістичного забезпечення комп'ютерних лінгвістичних систем тощо. В природних мовах є ситуації, коли явища, залежні від контексту, описано як не залежні від контексту, тобто в термінах контекстно-вільних граматик. При цьому опис ускладнений через утворення нових категорій і правил. В статті подано особливості процесу введення нових обмежень на класи цих граматик через введення нових правил. За кількості символів в правій частині правил не меншої ніж лівій отримано нескорочені граматики. Потім заміною лише одного символу отримано контекстно-залежні граматики. За наявності в лівій частині правила лише одного символу отримано контекстно-вільні граматики. Жодних наступних природних обмежень на ліві частини правил накласти вже не можна.

Застосування теорії породжувальних граматик для вирішення завдань прикладної та комп'ютерної лінгвістики на рівні морфології та синтаксису дає змогу створювати системи синтезу мови та текстів, а також підручники практичної морфології, таблиці словозміни, укладати списки морфем (афіксів, коренів), визначати продуктивності та частотності морфем, частоти реалізації в текстах різних граматичних категорій (категорій роду, відмінка, числа тощо) для конкретних мов. Розроблені на основі породжувальних граматик моделі використовують для забезпечення функціонування комп'ютерних лінгвістичних систем, призначених для аналітико-синтетичного опрацювання текстового контенту, в інформаційно-пошукових системах тощо. Корисно вводити все нові і нові обмеження на ці граматики, отримуючи вужчі їхні класи. Описуючи складне коло явищ, обмежують набір використовуваних засобів опису, розглядаючи і такі засоби, які подають у загальному випадку свідомо недостатніми. Дослідження починають із мінімальних засобів; кожного разу, коли їх недостатньо, поступово вводять (дрібнішими порціями) нові засоби; завдяки цьому вдається точно визначити, якими засобами можна/не можна обійтися для опису того або іншого явища для розуміння його природи.

Розглянуто відомі способи і підходи до вирішення проблеми автоматичного опрацювання текстового контенту та виділено недоліки й переваги існуючих підходів та результатів у галузі синтаксичних аспектів комп'ютерної лінгвістики. Сформовано загальні концептуальні принципи моделювання словозмінних процесів при утворенні текстових масивів на прикладі українських та німецьких речень, потім, запропонувавши синтаксичні моделі та словозмінні класифікації лексичного складу українських та німецьких речень, розроблено лексикографічні правила синтаксичного типу для автоматизованого опрацювання цих речень. За цією методикою можна досягти вищих показників надійності порівняно з відомими аналогами, а також високої ефективності прикладних застосувань при побудові нових інформаційних технологій лексокографування та дослідження словозмінних ефектів природних мов. Робота має практичну цінність, оскільки запропоновані моделі та правила дають змогу ефективно організувати процес створення лексикографічних систем опрацювання текстового контенту синтаксичного типу.

1. Английская грамматика в доступном изложении // Режим доступу: http://real-english.ru/crash/lesson3.htm. 2. Анісімов А. В. Алгоритмічна модель асоціативно-семантичного контекстного аналізу текстів природною мовою / А. В. Анісімов, О. О. Марченко, А. О. Никоненко // Пробл. програмув. – 2008. – № 2, 3. – С. 379–384. 3. Анисимов А. В. Компьютерная лингвистика для всех: мифы, алгоритмы, язык / А. В. Анисимов. – К.: Наукова думка, 1991. – 208 с. 4. Апресян Ю. Д. Идеи и методы современной структурной лингвистики / Ю. Д. Апресян. – М.: Просвещение, 1966. – 305 с. 5. Апресян Ю. Д. Непосредственно составляющих метод / Ю. Д. Апресян // Лингвистический энциклопедический словарь под ред. В. Н. Ярцевой. – М.: Советская энциклопедия, 1990. – Режим доступа: http://tapemark.narod.ru/les/332a.html. 6. Арсентьева Н.Г. О двух способах порождения предложений русского языка / Н. Г. Арсентьева // Проблемы кибернетики. – 1965. – Вып. 14. – С. 189–218. 7. Багмут А. Й. Порядок слів / А. Й. Багмут // Українська мова: Енцикл. – 3-тє вид., зі змінами і доп. – К.: В-во "Укр. енциклопедія" ім. М.П. Бажана, 2007. – С. 675-676. 8. Бильгаева Н. Ц. Теория алгоритмов, формальных языков, грамматик и автоматов: учеб. пособие / Н. Ц. Бильгаева. – Улан-Удэ: Изд-во ВСГТУ, 2000. – 51 с. 9. Большакова Е. И. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Е. И. Большакова, Э. С. Клышинский, Д. В. Ландэ, А. А. Носков, О. В. Пескова, Е. В. Ягунова. – М.: МИЭМ, 2011. – 272 с. 10. Висоцька В. А. Генерування речень українською за допомогою породжувальних граматик / В. А. Висоцька, Т. В. Шестакевич // Міжнародна наукова конференція "Інтелектуальні системи прийняття рішеньта проблеми обчислювального інтелекту (ISDMIT'2012)", Євпаторія. – 27–31 травня 2012. – С. 48–50. 11. Волкова И. А. Формальные грамматики и языки. Элементы теории трансляции / И .А. Волкова, Т. В. Руденко: учеб. пособие для студентов II курса – 2-е изд., перераб. и доп. – М.: Издательский отдел факультета вычислительной математики и кибернетики МГУ им. М.В.Ломоносова, 1999. – 62 с. 12. Гакман О. В. Генеративно-трансформаційна лінгвістика Н. Хомського як вираження його лінгвістичної філософії / О. В. Гакман // Мультиверсум.

Філософський альманах. – К.: Центр духовної культури, 2005. – № 45. – С. 98–114. 13. Герасимов А. С. Лекции по теории формальных языков / А. С. Герасимов. – Режим доступа: http://gas-teach.narod.ru/au/tfl/tfl01.pdf. 14. Гладкий А. В. Синтаксические структуры естественного языка в автоматизированных системах общения / А. В. Гладкий. – М.: Наука, 1985. – 144 с. 15. Гладкий А. В. Элементы математической лингвистики / А. В. Гладкий, И. А. Мельчук. – М.: Наука, 1969. – 192 с. 16. Гладкий А. В. Формальные грамматики и языки / А. В. Гладкий. – М.: Наука, 1973. – 368 с. 17. Гросс М. Теория формальных грамматик / М. Гросс, А. Лантен // Перевод с франц. И. А. Мельчука под редакцией А. В. Гладкого. – М.: Мир, 1971. – 294 с. 18. Дарчук Н. П. Комп'ютерна лінгвістика (автоматичне опрацювання тексту): підручник / Н. П. Дарчук. – К.: ВПЦ "Київський університет", 2008. – 351 с. 19. Демешко І. Типологія морфонологічних моделей у віддієслівному словотворенні сучасної української мови / І. Демешко // Збірник наукових праць "Лінгвістичні студії". Розділ V. Словотвір: напрями, аспекти дослідження. Морфонологія. – Донецьк, 2009. – № 19. – С. 162–167. 20. Зубков М. Українська мова: Універсальний довідник / М. Зубков. – К.: ВД "Школа", 2004. – 496 с. 21. Ингве В. Гипотеза глубины / В. Ингве // Новое в лингвистике. – М., 1965. – Вып. IV. – С. 126–138. 22. Любченко Т. П. Лексикографічні системи граматичного типу та їх застосування в засобах автоматизованого опрацювання мови: автореф. дис. канд. техн. наук: спец. 10.02.21 / Т. П. Любченко. – К., 2011. – 19 с. 23. Мартыненко Б. К. Языки и трансляции: учеб. пособие / Б. К. Мартыненко // 2-е изд., испр. и доп. – СПб.: Изд-во СПб. ун-та, 2008. – 257 с. 24. 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. – К., 2005. – 15 с. 25. Носков С. А. Самоучитель немецкого языка. Deutsch für sie / С. А. Носков. – К.: Наука, 1999. – 400 с. 26. Падучева Е. В. О связях глубины по Ингве со структурой дерева починений / Е. В. Падучева // Научно-техническая информация. – 1967. – № 6. – С. 38–43. 27. Партико З. В. Прикладна і комп'ютерна лінгвістика. Вступ до спеціальності: навч. посіб. / З. В. Партико. – Л.: Афіша, 2008. – 224 с. 28. Пентус А. Е. Теория формальных языков: учеб. пособие / А. Е. Пентус, М. Р. Пентус. – М.: Изд-во ЦПИ при механико-математическом ф-те МГУ, 2004. – 80 с. 29. Попов Э. В. Общение с ЭВМ на естественном языке / Э. В. Попов. – М.: Наука, 1982. – 360 с. 30. Постнікова О. М. Німецька мова. Розмовні теми: лексика, тексти, діалоги, вправи / О. М. Постнікова. – К.: А. С. К, 2001. – Т. 1. – 400 с. 31. Постнікова О. М. Німецька мова. Розмовні теми: лексика, тексти, діалоги, вправи / О. М. Постнікова. – К.: А.С.К, 2001. – Т. 2. – 320 с. 32. Потапова Г. М. Морфонологія віддієслівного словотворення (на матеріалі словотвірних гнізд з вершинами - дієсловами та віддієслівних словотвірних зон): Дис. канд. наук: 10.02.02 // Г. М. Потапова. – 2008. – 19 с. 33. Русаченко Н. П. Морфонологічні процеси у словозміні та словотворі староукраїнської мови другої половини XVI – XVIII ст.: автореф. дис. на здобуття наук. ступеня канд. філол. наук: спец. 10.02.01 / Н. П. Русаченко. – К., 2004. – 24 с. – Режим дос-тупу: http://auteur.corneille-moliere.com/?p=history&m=corneille_moliere&l=rus. 34. Торосян О. М. Функціональні характеристики прислівників міри та ступеня в сучасній англійській мові: автореф. дис. на здобуття наук. ступеня канд. філол. наук / О. М. Торосян. – Режим доступу: http://disser.com.ua/contents/6712.html. 35. Туришева О. О. Порушення рамкової конструкції в сучасній німецькій мові: функціональний аспект, нормативний статус: автореф. дис. канд. філол. наук: спец. 10.02.04 / О. О. Туришева. – Одеса, 2012. – 20 с. 36. Український правопис / Ін-т мовознавства ім. О. О. Потебні НАН України, Ін-т укр. мови НАН України. – К.: Наук. думка, 2007. – 288 с. 37. Фомичев В. С. Формальные языки, грамматики и автоматы / В. С. Фомичев. – Режим доступа: http://www.proklondike.com/books/thproch/. 38. Хомский Н. О некоторых формальных свойствах граматик / Н. Хомский // Кибернетический сборник. – М.: Мир, 1962. – № 5. – С. 279–311. 39. Хомский Н. Формальный анализ естественных языков / Н. Хомский, Дж. Миллер // Кибернетический сборник. – М.: Мир, 1965. – № 1. – С. 231–290. 40. Хомский Н. Язык и мышление / Н. Хомский // Публикации ОСиПЛ. Серия монографий. – М.: Издательство Московского университета, 1972. – № 2. – 122 с. 41. Хомский Н. Синтаксические структуры / Н. Хомский // Сборник «Новое в лингвистике». – М.: ИЛ, 1962. – № 2. – С. 412–527. 42. Чепурна З. В. Трансформація порядку слів у простому реченні при перекладі з німецької мови українською / З. В. Чепурна // Наукові записки, серія «Філологічні науки (мовознавство)»: у 5 ч. –

Кіровоград: РВВ КДПУ ім. В. Винниченка, 2010. – Вип. 89 (1). – С. 232–236. 43. Шаров С. А. Средства компьютерного представления лингвистической информации / С. А. Шаров. – Режим доступу : http://www.ksu.ru/eng/science/ittc/vol000/002/. 44. Шестакевич Т. В. Застосування породжувальних граматик для генерування речень українською мовою / Т. В. Шестакевич, В. А. Висоцька // Східно-Європейський журнал передових технологій. – Харків, 2012. – № 3/2 (57). – С. 51–53. 45. Шульжук К. Синтаксис української мови: Підручник / К. Шульжук. – К.: Академія, 2004. – 397 с. 46. Щербина Ю. М. Предмет математичної лінгвістики / Ю. М. Щербина // Вісник Нац. ун-ту "Львівська політехніка", серія «Інформаційні системи та мережі». – 2002. – № 464. – С. 340–349. 47. Щербина Ю. М. Науковий напрям та навчальна дисципліна "Математична лінгвістика" / Ю. М. Щербина, Т. В. Шестакевич, В. А. Висоцька // Вісник Нац. ун-ту "Львівська політехніка", серія «Інформаційні системи та мережі». – 2010. – № 673. – С. 384–392. 48. Шрейдер Ю. А. Характеристики сложности структуры текста / Ю. А. Шрейдер // Научно-техническая информация. – № 7. – 1966.– С. 34–41. 49. Chomsky N. Three models for the description of language / N. Chomsky. – I.R.E. Trans. PGIT 2, 1956. – P. 113–124. (Русский перевод: Хомский Н. Три модели для описания язика / Н. Хомский // Кибернетический сборник. – М.: ИЛ, 1961. – № 2. – С. 237–266). 50. Chomsky N. On certain formal properties of grammars, Information and Control 2 / N. Chomsky // A note on phrase structure grammars, Information andControl 2, 1959. – P. 137–267, 393–395. (Русский перевод: Хомский Н. Заметки о грамматиках непосредственных составляющих / Н. Хомский // Кибернетический сборник. – М.: ИЛ, 1962. – № 5. – С. 312–315). 51. Chomsky N. On the notion «Rule of Grammar» / N. Chomsky // Proc. Symp. Applied Math., 12. Amer. Math. Soc., 1961. (Русский перевод: Н. Хомский. О понятии «правило грамматики» / Н. Хомский // Сб. Новое в лингвистике. – М.: Прогресс, 1965. – № 4. – С. 34–65). 52. Chomsky N. Context-free grammars and pushdown storage / N. Chomsky // Quarterly Progress Reports, № 65, Research Laboratory of Electronics, M.I.T., 1962. 53. Chomsky N. Formal properties of grammars / N. Chomsky // Handbook of Mathemati-Mathematical Psychology, 2, ch. 12, Wiley, 1963. – P. 323–418. (Русский перевод: Н. Хомский. Формальные свойства грамматик / Н. Хомский // Кибернетический сборник. – М.: ИЛ, 1966. – № 2. – С. 121–230). 54. Chomsky N. The logical basis for linguistic theory / N. Chomsky // Proc. IX-th Int. Cong. Linguists, 1962. (Русский перевод: Хомский Н. Логические основы лингвистической теории / Н. Хомский // Сб. Новое в лингвистике. – М.: Прогресс, 1965. – № 4. – С. 465–575). 55. Chomsky N. Finite state languages / N. Chomsky, G.A. Miller // I nformation and Control 1, 1958. – P. 91–112. (Русский перевод: Хомский Н. Языки с конечным числом состояний. Кибернетический сборник. – М.: ИЛ, 1962. – № 4. – С. 231–255). 56. Chomsky N. Introduction to the formal analysis of natural languages / N. Chomsky, G. A. Miller // Handbook of Mathematical Psychology 2, Ch. 12, Wiley, 1963. – P. 269–322. (Русский перевод: Хомский Н. Введение в формальный анализ естественных языков / Н. Хомский, Д. Миллер // Кибернетический сборник. – М.: Мир, 1965. – № 1. – С. 229–290). 57. Chomsky N. The algebraic theory of context-free languages / N. Chomsky, M.P. Schützenberger // Computer programming and formal systems, North-Holland, MR152391. – Amsterdam, 1963. – P. 118–161. (Русский перевод: Хомский Н. Алгебраическая теория контекстно-свободных языков / Н. Хомский, М. Шютценберже // Кибернетический сборник, новая серия. – М.: Мир, 1966. – № 3. – С. 195–242). 58. Bar-Hillel Y. Finite state languages: formal representation and adequacy problems / Y.Bar-Hillel, E. Shamir //Bulletin of the Research Council of Israel. – 8F, № 3. – 1960. – P. 155–166. 59. Bobrow D.G. Syntactic analysis of English by computer – a survey / D. G. Bobrow // AFIPS conference proceedings. – 24, Baltimore – London. – 1963. – P. 365–387. 60. English Verbs (Part 1) – Basic Terms. – Режим доступу: http://sites.google.com/site/englishgrammarguide/Home/english-verbs--part-1----basic-terms. 61. Hays D. G. Automatic language data processing / D. G. Hays // Computer applications in behavioral sciences, Englewood Cliffs (N. J.) . – 1962. – P. 394–421. 62. Postal P.M. Limitations of phrase structure grammars / P.M. Postal // The structure of language. Readings in the philosophy of language, Englewood Cliffs (N. J.). – 1964. – P. 137–151. 63. Tesniere L. Elements de syntaxe structurale / L. Tesniere. – P. 1959. 64. Tosh L.W. Syntactic translation, The Hague / L. W. Tosh. – 1965. 65. Yngve V. H. A model and a hypothesis for language structure / V. H. Yngve // Proceedings of American phylosophical society. – 1960. – 104, № 5. – P. 444–466. 66. Yngve V. H. Random generation of English sentences / V.H. Yngve // Teddington (National physical laboratory. Paper 6). – 1961. 68. Varga D. Yngve's hypothesis and some problems of the mechanical analysis / D. Varga // Computational Linguistics. – III. – 1964. – P. 47–74.