КОМП'ЮТЕРНА ТА МАТЕМАТИЧНА ЛІНГВІСТИКА

УДК 81:004.93
В.А. Висоцька
Національний університет "Львівська політехніка",
кафедра інформаційних систем та мереж

ОСОБЛИВОСТІ ГЕНЕРУВАННЯ СЕМАНТИКИ РЕЧЕННЯ ПРИРОДНОЮ МОВОЮ ЗА ДОПОМОГОЮ ПОРОДЖУВАЛЬНИХ НЕОБМЕЖЕНИХ ТА КОНТЕКСТОЗАЛЕЖНИХ ГРАМАТИК

Описано застосування породжувальних граматик у лінгвістичному моделюванні. Опис моделювання синтаксису речення застосовують для автоматизації процесів аналізу та синтезу природномовних текстів.

Ключові слова: породжувальні граматики, структурна схема речення, комп'ютерна лінгвістична система.

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.

Key words: generative grammar, structured scheme sentences, computer linguistic system.

Вступ. Загальна постановка проблеми

На сучасному етапі розвитку потреба в розробленні загальних та спеціалізованих лінгвістичних систем змушує активно використовувати прикладну та комп'ютерну лінгвістику в галузі інформаційних технологій [3–4, 7–14, 16, 22, 25–28, 32–33, 35, 43, 50–51, 53–55, 66]. Розроблення математичних моделей мовлення для забезпечення комп'ютерних лінгвістичних систем дає змогу реалізовувати такі завдання прикладної лінгвістики, як аналіз/синтез усного/писемного текстового контенту, описування/індексування текстового контенту, перекладання текстів, створення лексикографічних баз даних тощо [7–8, 10–15, 25–28, 32–33, 43, 50–51, 53–55, 66]. Лінгвістичний аналіз текстового контенту складається з декількох послідовних процесів – графемного, морфологічного, синтаксичного та семантичного аналізу [2–5, 9–10, 16–23, 28–30, 34]. Для кожного з цих етапів створено відповідні моделі та алгоритми [2–5, 9–10, 15–21, 44–49, 56–64]. Ефективним інструментом лінгвістичного моделювання на синтаксичному та семантичному рівні мови є головна частина комбінаторної лінгвістики – теорія породжувальних граматик, початок якої закладений у роботах американського лінгвіста Н. Хомські [16–20, 45–48, 56–64]. Він використав прийом формального аналізу граматичної структури фраз для виділення синтаксичної структури (складових) як основної схеми фрази, незалежно від її значення [16]. Ідеї Н. Хомські розвинув радянський лінгвіст А.В. Гладкий [18–20], застосувавши поняття дерев залежності та систем складових для моделювання синтаксичного рівня мови [30, 50]. Він запропонував спосіб моделювання синтаксису за допомогою синтаксичних груп, що виділяють складові словосполучень як одиниці побудови дерева залежностей, – таке подання об'єднує переваги методу безпосередніх складових і дерев залежностей [4–5, 50].

Перевагами моделювання за допомогою породжувальних граматик є можливість опису не лише синтаксичного рівня мови (правила утворення речень зі словоформ) [12–14, 18–20, 48, 51], але й морфемного (правила утворення словоформ із морфів) [9–10, 23, 38–41, 55] та семантичного (правила утворення змістовних речень та текстів) [2, 29–30, 47, 49, 56]. Це використовують для

автоматизації процесів словозміни/словотворення, рубрикації або визначення ключових слів та формування дайджестів текстового контенту [7–8, 10–11, 23–28, 43, 66]. Наприклад, у разі використання автоматичного морфологічного синтезу комп'ютерна лінгвістична система утворює необхідні словоформи на основі заданих вимог до словоформ та баз даних морфем [32].

Зв'язок висвітленої проблеми із важливими науковими та практичними завданнями

Зважаючи на важливість забезпечення автоматичного опрацювання текстового контенту в сучасних інформаційних засобах (наприклад, інформаційно-пошукових системах, системах машинного перекладу, семантичного, статистичного, оптичного та акустичного аналізу і синтезу мови, автоматизованого редагування, екстракції знань з текстового контенту, реферування та анотування текстового контенту, індексування текстового контенту, навчально-дидактичних, менеджменту лінгвістичних корпусів, інструментальних засобах укладання словників різних типів тощо), фахівці інтенсивно шукають нові моделі, способи їх опису та методи автоматичного опрацювання текстового контенту [7–8, 10–11, 23–28, 43, 66]. Одним із таких способів є розроблення загальних принципів побудови лексикографічних систем синтаксичного типу та побудови за цими принципами зазначених систем опрацювання текстового контенту для конкретних мов [7–8, 10–11, 23–28, 43, 66].

Дослідники-мовознавці у галузі морфології, морфонології, в структурній лінгвістиці виділили різні структури для опису словоформ [7–8, 10–15, 25–28, 32–33, 43, 50–51, 53–55, 66]. З початком розвитку теорії породжувальних граматик лінгвісти зосередили увагу не лише на описі готових словоформ, а й на процесах їх синтезу [38]. В україністиці плідними є дослідження лінгвістів у функціональному аспекті [2–3, 6, 12–14, 22–24, 28, 30, 32–33, 39, 42–43, 51–55], зокрема, теоретичні проблеми морфонологічного опису [9–10, 23, 38–41, 55], питання класифікації морфемної і словотвірної структури дериватів української мови [23], закономірності комбінаторики афіксів [23, 55], моделювання словотвірного механізму сучасної української мови в словниках інтегрального типу [23, 38–39], принципи внутрішньої організації слова [23, 55], а також питання структурної організації відіменних дієслів і суфіксальних іменників із демінутивним значенням [23, 38, 55], проблеми словотвірної мотивації в процесі формування дериватів [23], закономірності реалізації морфонологічних явищ в українському словотворенні [23, 39, 55], морфонологічні модифікації у процесі словозміни дієслова [23, 38, 55], морфонологічні процеси при словозміні та словотворенні прикметників сучасної української літературної мови [23, 38–39], аналіз та опрацювання текстового контенту [7–8, 10–14, 22, 25–28, 30, 32–33, 43, 55, 66] тощо.

Такий динамічний підхід сучасної лінгвістики в аналізі морфологічного рівня мови із зосередженням уваги дослідника на розробленні морфологічних правил дає змогу ефективно застосовувати результати теоретичних досліджень на практиці – для побудови комп'ютерних лінгвістичних систем опрацювання текстового контенту різного призначення [7–8]. Одну із перших спроб застосувати для лінгвістичного моделювання теорію породжувальних граматик зробили А.В. Гладкий та І.А. Мельчук [18–20]. Напрацювання Н. Хомські [16–20, 45–48, 56–64] та А.В. Гладкого [18–20], дослідження М. Гросса і А. Лантена [21], А.В. Анісімова [2–3], Ю.Д. Апресяна [4–5], Н.Ц. Більгаєвої [9], І.А. Волкової та Т.В. Руденка [15], Є.І. Большакової, Е.С. Клишинського, Д.В. Ланде, А.А. Носкова, О.В. Пескової та Є.В. Ягунової [10–11, 26–27], А.С. Герасимова [17], Б.К. Мартиненко [29], А.Є. Пентуса та М.Р. Пентуса [34], Е.В. Попова [35], В.С. Фомічева [44] застосовні до розроблення таких засобів опрацювання текстового контенту, як інформаційно-пошукові системи, системи машинного перекладу, анотування текстового контенту, морфологічний, синтаксичний та семантичний аналіз текстового контенту, навчально-дидактичні системи опрацювання текстового контенту, до лінгвістичного забезпечення спеціалізованих програмних систем тощо [7–8, 10–14, 22, 25–28, 30, 32–33, 43, 55, 66].

Аналіз останніх досліджень і публікацій

Лінгвістичний аналіз контенту складається з трьох етапів: *морфологічний*, *синтаксичний* та *семантичний* [2–5, 10, 16–20, 45–48, 56–64]. Мета морфологічного аналізу полягає в здобутті основ

(словоформ без флексії) зі значеннями граматичних категорій (наприклад, частина мови, рід, число, відмінок) для кожної зі словоформ [10, 16–20, 45–48, 55–64]. Розрізняють *точні* та *наближені* методи морфологічного аналізу [10]. В точних методах використовують словники основ слів або словоформи, в наближених – експериментально встановлені зв'язки між фіксованими літеросполученнями словоформ та їх граматичним значенням [10, 55].

Використання словника словоформ у *точних методах* спрощує використання морфологічного аналізу. Наприклад, в українській мові вирішують проблему чергування голосних і приголосних при зміні умов вживання слова [10, 19, 55]. Тоді знаходження основ слів і граматичних ознак зводять до пошуку в словнику і вибору відповідних значень. А потім використовують морфологічний аналіз, якщо не знайдено шуканої словоформи в словнику. За достатньо повного тематичного словника швидкість опрацювання текстового контенту висока, але використання об'єму необхідної пам'яті в декілька разів більше, ніж у разі використання словника основ [7–8, 10, 19, 55]. Морфологічний аналіз з використанням словника основ базується на флективному аналізі та точному виділенні основи слова. Головна проблема тут пов'язана з омонімією основ слів. Для її усунення перевіряють сумісність виділеної основи слова і його флексії [10, 19, 45–48, 55–64].

Згідно з *наближеними методами* морфологічного аналізу визначають граматичний клас слова за кінцевими літерами і літеросполученнями [10]. Спочатку виділяють основу слова: від закінчення слова послідовно від'єднують по одній літері та отримані літеросполучення порівнюють із списком флексій з відповідного граматичного класу [10, 19, 55]. Якщо отримано збіг, остаточну частину слова визначають як її основу [10]. Під час проведення морфологічного аналізу виникають неоднозначності визначення граматичної інформації, які знімаються після синтаксичного аналізу [10, 12–14, 19, 56–64]. Завданням синтаксичного аналізу є здійснення граматичного розбору речень на підставі даних зі словника [10, 19, 55]. На цьому етапі виділяють підмет, присудок, прикметник тощо, між якими вказують зв'язки у вигляді дерева залежностей [10, 12–14, 19, 33, 55].

Будь-які засоби синтаксичного аналізу складаються з двох частин: бази знань про конкретну природну мову й алгоритму синтаксичного аналізу, тобто набору стандартних операторів опрацювання текстового контенту на основі цих знань [7–8, 10, 12–14, 19, 55, 66]. Джерелом граматичних знань є дані з морфологічного аналізу та різні заповнені таблиці понять та лінгвістичних одиниць. Це результат емпіричного опрацювання текстового контенту природною мовою експертами з метою виділення основних закономірностей для проведення синтаксичного аналізу. Основу таблиць лінгвістичних одиниць становлять сукупності конфігурацій або набори валентностей (синтаксичних і семантико-синтаксичних залежностей). Це є списки/словники лексичних одиниць із вказівкою для кожної з них всіх можливих варіантів зв'язків із іншими одиницями виразу природною мовою [7–8, 10, 19, 55, 66]. Виконуючи синтаксичний аналіз, необхідно досягти повної незалежності правил перетворення даних таблиць від їх вмісту, щоб зміна цього вмісту не вимагала перебудови алгоритму [10, 12–14, 19, 66].

Словник V складається зі скінченної непорожньої множини лексичних одиниць [65]. 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, який повинні задовольняти лексичні одиниці, щоб належати мові [16–20, 45–48, 56–64]. Ще один важливий спосіб задати мову – через використання породжувальної граматики. 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.

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 $x \to h$, де x та h – ланцюжки над V. Множину $V \setminus T$

позначають *N*, її елементи є *нетермінальними* (*допоміжними*) лексичними одиницями, *не терміналами* [18-20]. Граматики класифікують за типами продукцій, на які накладено певні обмеження (табл.1) [18–20, 45–48, 56–64].

Таблиця 1

Класифікація граматик за типами продукцій

Граматика	Тип	Опис
G_0	Необмежена	де x – довільний ланцюжок, що містить хоча б один нетермінальний символ, h – довільний ланцюжок над V.
G_1	Контекстно-залежна	В множині продукцій P є продукція вигляду $gxd \to ghd$, $\lvert x \rvert \le \lvert h \rvert$ (але не у формі $x \to h$), тому x можна замінити на v лише в оточенні ланцюжків g і d, тобто у відповідному контексті.
G_2	Контекстно-вільна	Нетермінал A у лівій частині продукції $A \to h$ може бути замінений ланцюжком h у довільному оточенні щоразу, коли він трапляється, тобто незалежно від контексту.
G_3	Регулярна	Можуть бути лише продукції $A \to aB$, $A \to a$, $S \to l$, де A, B – нетермінали, a – термінал, l – порожній ланцюжок.

Термінальні лексичні одиниці є словоформами природної мови, нетермінальні лексичні одиниці – синтаксичні категорії, а термінальні ланцюжки, що виводяться, – правильні вирази цієї мови [18–20, 45–48, 56–64]. Тоді виведення виразу природно інтерпретують як його синтаксичну структуру, яка подана в термінах породжувальної граматики [12–14, 51]. Множина виразів природною мовою володіє низкою 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 цієї 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.

Граматика G_0 не задовольняє вказане обмеження – в ній є продукції, що скорочують ланцюжки [18–20, 45–48, 56–64]. Однак мова $L(G_0)$ є рекурсивною. Мови, породжені нескорочувальними граматиками, легкорозпізнавані. У контекстозалежній граматиці G_1 є продукція $gAd \to ghd$, у якій хоча б один з ланцюжків g, d відмінний від Λ, а нетермінал A замінюють ланцюжком h лише в оточенні g та d, тобто у відповідному контексті. Мова є контекстозалежною, якщо існує принаймні одна контекстозалежна граматика, яка породжує цю мову [18–20, 45–48, 56–64].

Термін *правила утворення* запозичений з математичної логіки, де він позначає правила побудови правильних формул. В логіці розглядають інший тип правил – *правила перетворення*. Вони задають певні співвідношення між правильними формулами. Правила перетворення необхідні й для опису природних мов. Введення правил перетворення означає перехід до вищого рівня розгляду мови, а саме до семантичного рівня. Володіння мовою обов'язково передбачає уміння не лише побудувати правильну фразу, але і перейти від однієї фрази до інших, або повністю синонімічних їй, або що відрізняються від неї за сенсом на певну *величину*, наприклад, зробити зі ствердного речення питальне або негативне, з активної конструкції пасивну, змінити стилістичне забарвлення тексту, виразити ту саму думку різними способами тощо. Ці можливості не можна викласти в термінах граматик, і тому постає питання про розроблення формального апарату для правил перетворення стосовно природних мов. Відповідне завдання вперше чітко сформульовано в роботах Н. Хомські [18–20, 45–48, 56–64]. Висунута ним концепція швидко набула широкої популярності під назвою *трансформаційної граматики*: введення семантичного рівня опису мови.

Насправді, інваріантом всіх трансформацій зазвичай є сенс, тобто трансформації – це перетворення, що зберігають сенс. Отже, теорія трансформацій є теорією синонімії в мові. Опис синонімії повинен займати в лінгвістиці одне з центральних місць. Звідси випливає і первинна роль трансформацій. Проте трансформації належать не до того самого рівня, що граматики G_0 : граматики G_0 належать до синтаксичного рівня, а трансформації – до семантичного. Тобто недостатність граматики G_0 для опису мови є правильною в сенсі неохоплення граматики G_0 семантичного рівня. На синтаксичному рівні граматики G_0 є принципово цілком достатніми. Граматики, що породжують, розглядають в межах формальної теорії. Для трансформацій рівень формалізації не досягнутий: трансформаційні правила не сформульовані в термінах однієї простої операції. Завдання подальшої формалізації трансформацій є вельми актуальним. У роботах Н. Хомські [18–20, 45–48, 56–64] та деяких інших авторів [2–5, 9–10, 12–23, 32–35, 44, 50–51, 53–55] термін "граматика, що породжує", використовують в двох сенсах: у широкому – для позначення будь-якої системи формальних правил, що описують мову, з введенням трансформаційного і морфонологічного компонентів, і у вузькому – для позначення саме граматик. Цей термін завжди вживають у вузькому сенсі, за такого слововживання трансформаційні правила за межами граматики, що породжує.

Формулювання мети

Подання синтаксичної структури в термінах породжувальної граматики часто застосовують в лінгвістиці та багато разів досліджували в найрізноманітніших аспектах. Воно завоювало право на існування як в теоретичному плані, так і в роботах експериментального характеру (автоматичний переклад або реферування тощо). Граматики, у разі породження термінальних ланцюжків, наприклад, виразів природної мови, одночасно дають їх структуру. Необмежена граматика, що не укорочує, вже не володіє властивістю порівняння виразів з їх контекстозалежною структурою. В такій граматиці кожного разу замінюють не одну лексичну одиницю, а цілу їх групу. У виведенні неможливо однозначно вказати для кожної лексичної одиниці її предка, і тому виведення не перетворюється на контекстозалежну структуру. Лінгвістичне забезпечення використовують в багатьох інформаційних системах. Вдосконалення спілкування *машина-людина* є важливим актуальним завданням, яке вирішують через аналіз та синтез текстів на лінгвістичному рівні. З цією метою розглянемо процес лінгвістичного моделювання фраз природною мовою за допомогою породжувальних граматик. Для цього здійснимо належний опис словозмінних лінгвістичних систем – на основі лінгвістичного аналізу визначимо перелік відповідних лексичних одиниць, а також з'ясуємо систему правил, за допомогою яких одержують будь-які правильні форми виразів природною мовою, не отримавши при цьому жодного неправильного. В межах статті покажемо способи застосування апарату породжувальних граматик до моделювання синтаксису речень для різних мов – німецької та української. Для цього розберемо синтаксичну структуру речень, продемонструємо особливості процесу синтезу речень зазначених мов. Розглянемо вплив норм та правил мови на хід побудови граматик [12–14, 51].

Аналіз отриманих наукових результатів

Текстовий контент (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 текстового контенту є лінгвістичний аналіз змісту (наприклад, коментарі, форуми, статті тощо). Процес опрацювання текстового контенту поділяє контент на лексеми за допомогою кінцевих автоматів лінгвістичного аналізу текстів природної мови (рис. 1).

Як функціонально-семантико-структурна єдність текстовий контент володіє правилами побудови, виявляє закономірності змістового та формального з'єднання складових одиниць. Зв'язність текстового контенту проявляється через зовнішні структурні показники та формальну залежність компонентів тексту, а цілісність текстового контенту – через тематичну, концептуальну

та модальну залежність текстової інформації. Цілісність текстового контенту веде до змістової та комунікативної організації тексту, а зв'язність текстового контенту – до форми, структурної організації текстової інформації.

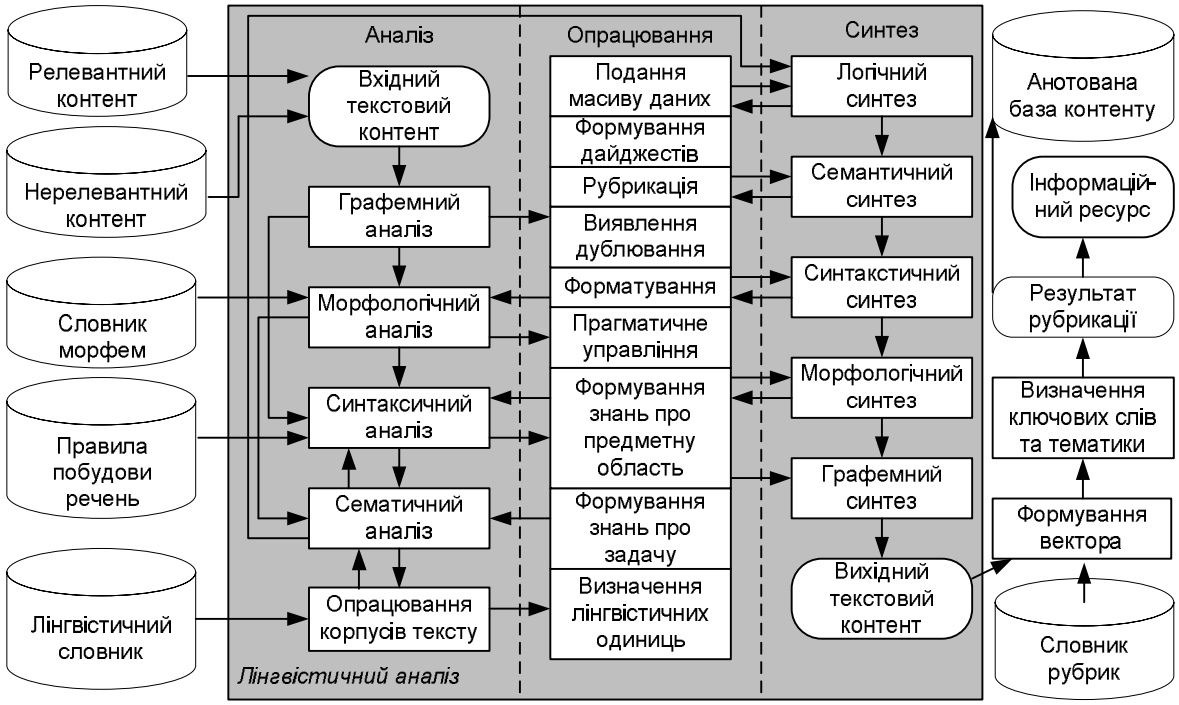


Рис. 1. Структурна схема лінгвістичного аналізу текстового контенту під час формування контенту

Розглянемо нескорочувальну граматику G з p лінгвістичними одиницями та ланцюжок x довжини n із термінальних лінгвістичних одиниць цієї граматики. Мова $L(G)$ є легкорозпізнаваною множиною за алгоритмом (алг. 1), який будує один за іншим будь-які виведення в граматиці G із початкового символу *S*. Кількість вживань продукцій у виведенні є його довжиною.

Алгоритм 1. Розпізнавання мови L(G).

Етап 1. Почергово до *S* застосовують продукції граматики G.

Етап 2. Перевірка побудованого ланцюжка x.

Крок 1. Якщо так, то перехід до етапу 3.

Крок 2. Якщо ні, то перехід до етапу 1.

Етап 3. Виведення ланцюжка x із *S*.

За цим алгоритмом процес виведення може бути нескінченним. Щоб уникнути цього, алгоритму задають скінченну множину виведень M (алг. 2).

Алгоритм 2. Розпізнавання мови L(G) з використанням множини виведень M.

Етап 1. Аналіз виведень ланцюжка x довжини n із початкового символу *S* граматики G.

Крок 1. Розрахувати кількість ланцюжків від *S* до x (жоден ланцюжок не повторюється). Оскільки граматика G є нескорочувальною, то жодний з ланцюжків цієї послідовності не довший за ланцюжок x (довжина $\le n$). Кількість різних ланцюжків довжини $\le n$ із p лексичних одиниць $\le p^n + p^{n-1} + p^{n-2} + ... + p^2 + p^1 + p^0 = \frac{p^{n+1}-1}{p-1} < p^{n+1}$, якщо $p > 1$ (кількість ланцюжків довжини n із p лексичних одиниць дорівнює p^n, довжини $(n-1)$ із p дорівнює p^{n-1} тощо; кількість ланцюжків із 0 символів становить $p^0 = 1$). Якщо $p^{n+1} = P$, різних ланцюжків таких послідовностей не більше за $P! + C_P^1 \cdot (P-1)! + C_P^2 \cdot (P-2)! + ... + C_P^{P-2} \cdot 2! + C_P^{P-1} \cdot 1!$. Тут у сумі P доданків її k-й

доданок дорівнює $C_P^k \cdot (P-k)! = \frac{P!}{k!(P-k)!} \cdot (P-k)! = \frac{P!}{k!} \le P!$, де $C_P^k = \frac{P!}{k!(P-k)!}$. Вся сума не більша за $P! \cdot P < (P+1)! = \left(p^{n+1}+1\right)! < \left(p^{n+2}\right)!$.

Крок 2. Сформувати послідовність ланцюжків від S до x. Із отриманих $\left(p^{n+2}\right)!$ послідовностей виведень в граматиці G утворюють шукану множину M.

Етап 2. Побудова множини виведень M за довільним ланцюжком x.

Етап 3. Перевірка отриманої скінченної множини виведень M.

Крок 1. Перебрати множину M для виявлення необхідного виведення або доведення, що такого виведення не існує. Якщо виведення в M не закінчується ланцюжком x, то перехід до кроку 3.

Крок 2. Якщо виведення в M закінчується ланцюжком x, то перехід до етапу 4.

Крок 3. Якщо кінець множини виведень M, то перехід до етапу 5, інакше перехід до етапу 3.

Етап 4. Формулювання позитивної відповіді: x виводиться з S. Перехід до етапу 5.

Етап 5. Формулювання негативної відповіді: x не виводиться з S.

Етап 6. Виведення результату.

Кількість кроків формування множини M для знаходження необхідного виведення не перевищує $\left(p^{n+2}\right)!$ та є великою для природних мов. Це вимагає для виконання такого алгоритму багато ресурсів та великої потужності. Тому необхідно вибрати множину, в якій кількість кроків під час розпізнавання перебуває у вказаній залежності від довжини ланцюжка, та виділити в класі рекурсивної множини достатньо вузький підклас. А за умови, що множина виразів є нескінченною, правила опрацювання їх є однотипнішими, що дає змогу розкрити суттєві закономірності виведення.

Для точнішого виведення ланцюжка x із S необхідно ввести ще додаткове обмеження: в кожній продукції $X \to Y$ ліва частина (X) має вигляд Z_1CZ_2 (C – одна лексична одиниця), а права частина (Y) – вигляд Z_1WZ_2 (W – непорожній ланцюжок). Тоді на кожному кроці виведення дозволено замінювати лише одну лексичну одиницю. Для будь-якої нескорочувальної граматики можна побудувати еквівалентну їй контекстозалежну граматику, наприклад,

$$P_1 = \{AB \to BA\} \approx P_2\{AB \to 1B, 1B \to 12, 12 \to B2, B2 \to BA\}$$

Послідовне вживання цих правил рівнозначне вживанню правила $AB \to BA$, причому заміна їх останнім не приведе до появи зайвих виведень, оскільки лексичні одиниці 1 і 2 є новими. В правилах граматики G замінюють лише одну лексичну одиницю (C). Ліва частина продукції (X) не обов'язково складається лише з цієї лексичної одиниці. Навколо C можуть бути інші лексичні одиниці (контекст), тобто $X = Z_1CZ_2$. Тоді продукція вигляду $Z_1CZ_2 \to Z_1WZ_2$ означає дозвіл замінювати C на W лише в контексті $Z_1...Z_2$ без зміни контексту та його розташування.

Ввівши нове обмеження (права частина будь-якої продукції містить не більше від двох лексичних одиниць), утворюють новий клас контекстовільних граматик, де в синтаксичних структурах виразів під час побудови дерева із кожної вершини виходить не більше від двох гілок. Тобто вираз поділяють завжди на дві половини (наприклад, *іменна група + дієслівна група*), кожну з цих половин знову поділяють навпіл тощо. Але бінарне подання виразів природної мови не завжди є задовільним та природним з погляду змістової лінгвістичної інтерпретації. Критерії вибору відповідного опису лежать поза теорією: цей вибір роблять на основі міркувань, що стосуються конкретних цілей і характеру поставленого завдання. Оскільки кількість лексичних одиниць в правій частині продукції вже є мінімальною, накладають обмеження на характер лексичних одиниць, які заміняють (якщо права частина кожної продукції складалася з однієї лексичної одиниці або має вигляд bB, де b – термінальна лексична одиниця, а B – синтаксична категорія). Це обмеження уточнює ланцюжок виведення, але вимагає великої потужності для обчислень.

Граматики володіють такою важливою в лінгвістичному аспекті властивістю. Тлумачитимемо термінальні символи як словоформи (деякої природної мови), допоміжні символи – як синтаксичні категорії (наприклад, V – дієслово, S – іменник, A – прикметник, $\overset{\vee}{V}$ – група дієслова, $\overset{\vee}{S}$ – група іменника), початковий символ – як R (речення), а термінальні ланцюжки, що виводяться, – як правильні речення цієї мови. Тоді виведення речення природно інтерпретують як його синтаксичну структуру, подану в термінах безпосередніх складових. Пояснимо сказане прикладами.

1. Весела посмішка твого сина наповнює мене безмежним щастям [12–14, 51].
2. In seinem bedeutendsten Werk zeigt er die bunte welt des ukrainischen Dorfes in ihrem einmaligen Reiz [31].

Побудуємо граматику G_1, яка породжуватиме фрази відповідної мови (української, англійської або німецької), синтаксично однотипні й дуже прості [1, 6, 10–14, 22, 24, 28, 31, 36–42, 49, 51–55, 65]. Випишемо лише схему цієї граматики; її термінальними символами є словоформи відповідної мови, а допоміжний словник містить вищеназвані синтаксичні категорії. Символи цих категорій забезпечені індексами, відповідними їх морфологічним ознакам, наприклад $S_{ж,од,р}$. Початковий символ позначається через R.

Схема граматики G_1. Зміст використовуваних позначень G_1 пояснено після табл. 2.

Таблиця 2

Правила формулювання речень українською мовою

№	Назва правила	Правило
I	Вибір структури R	$R \to \# \overset{\vee}{S}_{x,y,н,w} \overset{\vee}{V}_{y,тепер,w} \# .$
II	Розгортання іменної групи	1) $\overset{\vee}{S}_{x,y,z,3} \to \overset{\vee}{S}_{x,y,z,3} \overset{\vee}{S}_{x',y',p,w}$; 2) $\overset{\vee}{S}_{x,y,z,3} \to A_{x,y,z} \overset{\vee}{S}_{x,y,z,3}$; 3) $K_1 \overset{\vee}{S}_{x,y,z,w} K_2 \to K_1 S^{займ}_{x,y,z,w} K_2$, де K_1 – символ, відмінний від символу $A_{x,y,z}$, а K_2 – символ, відмінний від символу з індексом $z' = p$. Символи K_1 і K_2 є контекстними обмеженнями. Змістовий сенс їх введення в це правило: головний член іменної групи не повинен реалізовуватися особистим займенником, якщо йому передує визначення, виражене погодженим прикметником, або якщо за ним слідує іменна група в родовому відмінку, наприклад, неможливість *новий я* або *він ніжності*. У поетичній мові подібні поєднання припускають; 4) $\overset{\vee}{S}_{x,y,z,3} \to S_{x,y,z}$.
III	Розгортання дієслівної групи	1) $\overset{\vee}{V}_{y,тепер,w} \to V_{y,тепер,w} \overset{\vee}{S}_{x',y',зн,w'} \overset{\vee}{S}_{x'',y'',ор,w''}$; 2) $\overset{\vee}{V}_{y,тепер,w} \to V_{y,тепер,w} \overset{\vee}{S}_{x',y',ор,w'} \overset{\vee}{S}_{x'',y'',зн,w''}$; 3) $\overset{\vee}{V}_{y,тепер,w} \to V_{y,тепер,w} \overset{\vee}{S}_{x',y',зн,w'}$; 4) $\overset{\vee}{V}_{y,тепер,w} \to V_{y,тепер,w} \overset{\vee}{S}_{x',y',ор,w'}$.
IV	Реалізація синтаксичних категорій словоформами	1) $S_{ч,y,z} \to син_{y,z},...$; 2) $S_{ж,y,z} \to посмішка_{y,z},...$; 3) $S_{сер,y,z} \to щастя_{y,z},...$ 4) $S^{займ}_{x,од,z,1} \to я_z$; 5) $S^{займ}_{x,од,z,2} \to ти_z$; 6) $A_{x,y,z} \to веселий_{x,y,z}, безмежний_{x,y,z}, мій_{x,y,z}, твій_{x,y,z},...$; 7) $V_{y,тепер,w} \to наповнити_{y,тепер,w},...$

Кожен рядок у цій схемі є не одним правилом, а скороченим записом декількох правил .Так, рядок II.1 формує 648 правил формування словосполучень в українській мові [6, 12-14, 24, 42, 51, 52, 55]: $\overset{\vee}{S}_{ч,од,н,3} \to \overset{\vee}{S}_{ч,од,н,3} \overset{\vee}{S}_{ч,од,р,1}$; $\overset{\vee}{S}_{ч,од,р,3} \to \overset{\vee}{S}_{ч,од,р,3} \overset{\vee}{S}_{ч,од,р,1}$; ...; $\overset{\vee}{S}_{сер,мн,м,3} \to \overset{\vee}{S}_{сер,мн,м,3} \overset{\vee}{S}_{сер,мн,род,3}$, де скорочення *ч* – чоловічий рід, *од* – однина, *н* – називний відмінник, *р* – родовий відмінник, *сер* – середній рід, *мн* – множина, *м* – місцевий відмінник, 1 – перша особа іменника, 3 – третя особа іменника. Такий самий спосіб скорочення застосовують і в наступних прикладах. Однак для простоти формулювань називатимемо рядки таких скорочених записів *правилами* (табл. 2). У правилах IV не враховано узгодження A з іменником істот S у знахідному відмінку. Позначення:

– символ межі речення, який є термінальним (у тексті ліву межу реалізують великою літерою першого слова, а праву – крапкою); x, y, z, w – змінні характеристик словоформ, відповідні роду, числу, відмінку, особі, наприклад, $веселий_{ж,од,н} = весела$. Приклад виведення в граматиці G_1 для генерування українського речення подано на рис. 2.

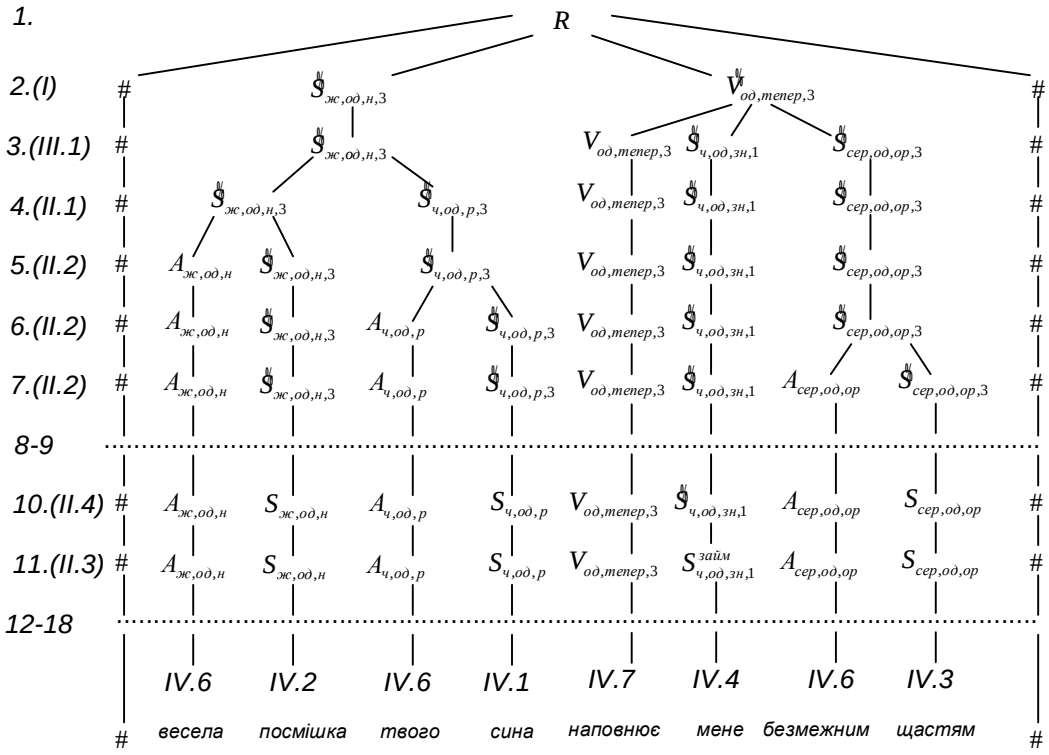


Рис. 2. Приклад граматики із фразовою структурою (тип 0) для генерування українського речення

Граматика G_1 здатна породжувати й інші фрази (які не обов'язково є змістовими), наприклад: *# я наповнюю тебе щастям #*, *# веселе безмежне моє щастя наповнює тобою твоїй посмішці щастя твоєї посмішки щастя #* тощо. Граматика G_1 породжує нескінченно багато різних фраз (на відміну від граматики G_0), оскільки до її складу входять так звані циклічні правила (II.1 і II.2). Особливість такого правила полягає в тому, що результат його застосування містить входження його лівої частини, так що воно завжди може бути застосоване до свого власного результату, що і приводить до нескінченної кількості фраз: так, поряд з групою *весела посмішка* можна отримати *весела весела посмішка*, далі *весела весела весела посмішка* тощо, тобто прикметник *весела* може бути повторений скільки завгодно разів. У зв'язку з цим постає питання про нескінченність кількості фраз в природній мові, відносно якої зазначимо, що в кожен певний момент кількість слів будь-якої природної мови скінченна. Крім того, максимальна довжина фраз, що трапляються в мові, практично обмежена: навряд чи люди вживають фрази більш ніж, скажімо, із 1000 слів. Звідси випливає, що кількість фраз в природній мові має бути скінченна. Але *вказати щонайдовшу фразу неможливо*: яку б фразу не запропонувати, завжди можемо подовжити її, додавши до неї, наприклад, ще один однорідний член або підрядне речення з *який*. В природній мові існують принципові можливості для побудови як завгодно довгих фраз, тобто потенційно реальні фрази будь-якої довжини, хоча на практиці великі фрази не використовують. Ця *потенційна необмеженість* довжини фраз не може не враховуватися формальними граматиками, оскільки їх завданням є моделювання принципових можливостей природної мови. Якщо ж довжини фраз, що породжуються граматикою, необмежені, то множина всіх цих фраз нескінченна.

Для німецької мови запис II.1 формує 288 правил: $\mathbb{S}_{m,einz,nom,3} \to \mathbb{S}_{m,einz,nom,3}\, \mathbb{S}_{m,einz,gen,1}$; $\mathbb{S}_{m,einz,gen,3} \to \mathbb{S}_{m,einz,gen,3}\, \mathbb{S}_{m,einz,gen,1}$; ...; $\mathbb{S}_{n,plur,akk,3} \to \mathbb{S}_{n,plur,akk,3}\, \mathbb{S}_{n,plur,akk,3}$, де m – Maskulinum (чоловічий рід), *einz* – Einzahl (однина), *nom* – Nominativ (називний відмінник), *gen* – Genitiv (родовий відмінник), *n* – Neutrum (середній рід), *akk* – Akkusativ (знахідний відмінник), *plur* – Plural (множина), 1 – erste Person Substantiv (перша особа іменника), 3 – dritte Person Substantiv (третя особа іменника) [12–13, 31, 36–37, 41, 49]. В табл. 3 подані правила формулювання речень німецькою мовою.

Приклад виведення в граматиці G_1 для генерування німецького речення подано на рис. 3.

Таблиця 3

Правила формулювання речень німецькою мовою

№	Назва правила	Правило
I	Вибір структури R	1) $R \to \#\mathbb{S}_{x,y,nom,w}\, \mathbb{V}_{y,pr,w}\, \#$; 2) $R \to \#\mathbb{S}_{x,y,dat,w}\, \mathbb{V}_{y,pr,w}\, \#$.
II	Розгортання іменної групи	1) $\mathbb{S}_{x,y,z,3} \to \mathbb{S}_{x,y,z,3}\, \mathbb{S}_{x',y',gen,w}$; 2) $\mathbb{S}_{x,y,z,3} \to A_{x,y,z}\, \mathbb{S}_{x,y,z,3}$; 3) $K_1 \mathbb{S}_{x,y,z,w}\, K_2 \to K_1 S^{pron}_{x,y,z,w} K_2$, де K_1 — символ, відмінний від символу, а K_2 — символ, відмінний від символу з індексом $z' = gen$. Символи K_1 і K_2 є контекстними обмеженнями. Змістовий сенс їх введення в це правило полягає в тому, що головний член іменної групи не повинен реалізовуватися особистим займенником, якщо йому передує визначення, виражене погодженим прикметником, або якщо за ним слідує іменна група в родовому відмінку, наприклад, неможливість *neu ich* (*новий я*) або *er Zärtlichkeit* (*він ніжності*). У поетичній мові подібні поєднання припускають; 4) $\mathbb{S}_{x,y,z,3} \to S_{x,y,z}$; 5) $\mathbb{S}_{x,y,z,3} \to \mathbb{S}_{x,y,z,3}\, \mathbb{S}_{x',y',akk,w}$; 6) $\mathbb{S}_{x,y,z,3} \to \mathbb{S}_{x,y,z,3}\, \mathbb{S}_{x',y',dat,w}$
III	Розгортання дієслівної групи	1) $\mathbb{V}_{y,pr,w} \to V_{y,pr,w}\, \mathbb{S}_{x',y',akk,w'}\, \mathbb{S}_{x'',y'',gen,w''}$; 2) $\mathbb{V}_{y,pr,w} \to V_{y,pr,w}\, \mathbb{S}_{x',y',gen,w'}\, \mathbb{S}_{x'',y'',akk,w''}$; 3) $\mathbb{V}_{y,pr,w} \to V_{y,pr,w}\, \mathbb{S}_{x',y',akk,w'}$; 4) $\mathbb{V}_{y,pr,w} \to V_{y,pr,w}\, \mathbb{S}_{x',y',akk,w'}\, \mathbb{S}_{x'',y'',gen,w''} \mathbb{S}_{x''',y''',dat,w'''}$; 5) $\mathbb{V}_{y,pr,w} \to V_{y,pr,w}\, \mathbb{S}_{x',y',akk,w'}\, \mathbb{S}_{x'',y'',dat,w''}$; 6) $\mathbb{V}_{y,pr,w} \to V_{y,pr,w}\, \mathbb{S}_{x',y',dat,w'}\, \mathbb{S}_{x'',y'',akk,w''}$; 7) $\mathbb{V}_{y,pr,w} \to V_{y,pr,w}\, \mathbb{S}_{x',y',dat,w'}\, \mathbb{S}_{x'',y'',akk,w''} \mathbb{S}_{x''',y''',dat,w'''}$; 8) $\mathbb{V}_{y,pr,w} \to V_{y,pr,w}\, \mathbb{S}_{x',y',dat,w'}\, \mathbb{S}_{x'',y'',dat,w''} \mathbb{S}_{x''',y''',akk,w'''}$; 9) $\mathbb{V}_{y,pr,w} \to V_{y,pr,w}\, \mathbb{S}_{x',y',akk,w'}\, \mathbb{S}_{x'',y'',dat,w''} \mathbb{S}_{x''',y''',dat,w'''}$; 10) $\mathbb{V}_{y,pr,w} \to V_{y,pr,w}\, \mathbb{S}_{x',y',nom,w'}$;
IV	Реалізація синтаксичних категорій словоформами	1) $S_{m,y,z} \to Reiz_{y,z}, \ldots$ (де *der Reiz* – *привабливість*); 2) $S_{f,y,z} \to Welt_{y,z}, \ldots$ (де *die Welt* – *світ*); 3) $S_{n,y,z} \to Werk_{y,z}, Dorf_{y,z}, \ldots$ (де *das Werk* – *робота*, *das Dorf* – *село*); 4) $S^{pron}_{x,einz,z,1} \to er_z$ (де *er* – *він*); 5) $S^{pron}_{x,plur,z,3} \to sie_z$ (де *sie* – *вони*); 6) $A_{x,y,z} \to sein_{x,y,z}, ihr_{x,y,z}, bedeutend_{x,y,z}, bunt_{x,y,z}, ukrainisch_{x,y,z}, einmaling_{x,y,z}, \ldots$ (де *sein* – *свій*, *ihr* – *їхній*, *bedeutend* – *важливий*, *bunt* – *барвистий*, *ukrainisch* – *український*, *einmalig* – *неповторний*); 7) $V_{y,pr,w} \to zeigen_{y,pr,w}, \ldots$ (де *zeigen* – *показати*).

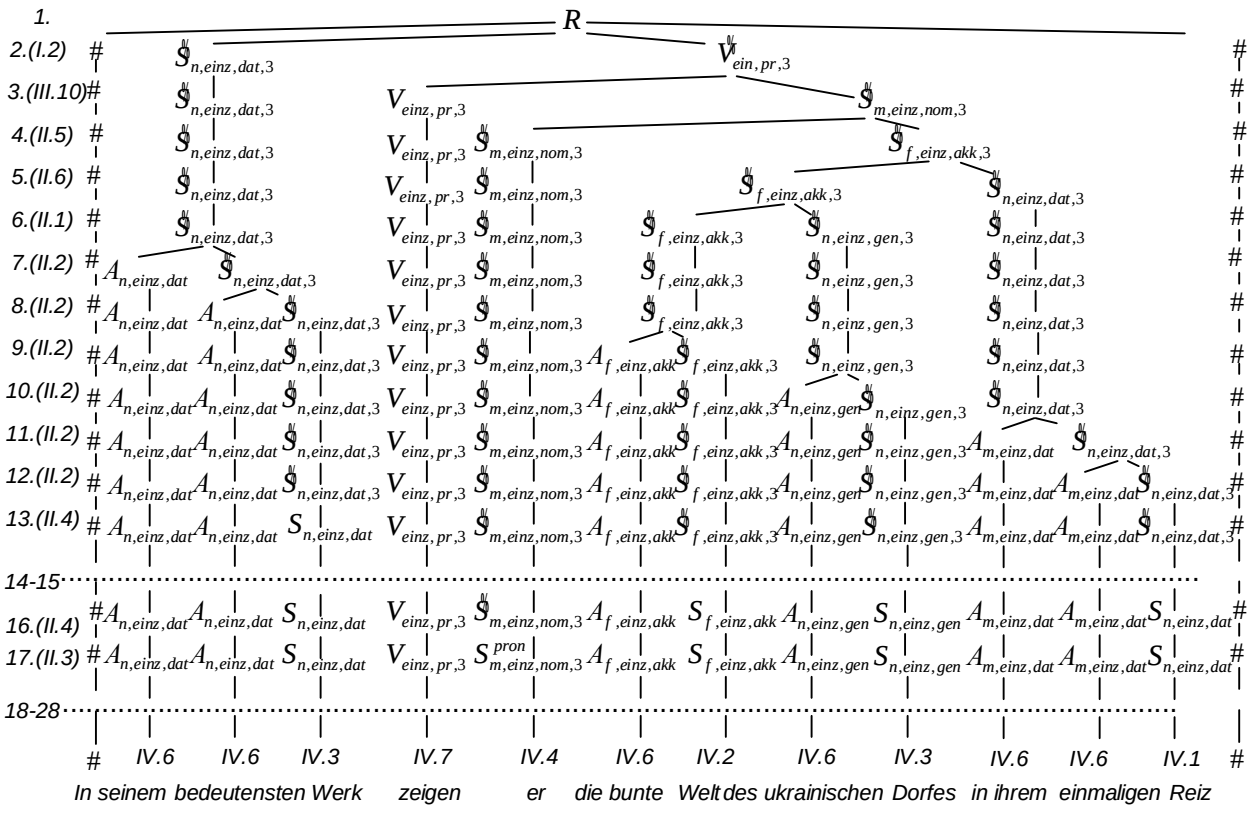


Рис. 3. Приклад граматики із фразовою структурою (тип 0) для генерування німецького речення

Повернемося до рис. 1, де кожен крок виведення полягає в розгортанні одного з символів попереднього ланцюжка (наприклад, з переходом від 2 до 3 ланцюжка символ $V_{од,тепер,3}$ розгортається в три символи – $V_{од,тепер,3}$ $S_{ч,од,зн,1}$ $S_{сер,од,ор,3}$), або в заміні його іншим (наприклад, з переходом від 10 до 11 символ $S_{ч,од,зн,1}$ замінюється на $S^{займ}_{ч,од,зн,1}$), інші ж символи переписують без зміни (правила підстановки). Розгортані, замінювані або переписувані символи є *предками*, а символи, які отримуємо в результаті розгортання, заміни або переписування, – їх *нащадками* (нащадки нащадків також є нащадками). З'єднаємо предків лініями з їх безпосередніми нащадками. Тоді отримаємо дерево складових, або синтаксичну структуру фрази в термінах безпосередніх складових (тип 1). Для ілюстрації цього явища вилучимо із схеми на рис. 1 всі символи-нащадки, що переписують без зміни (наприклад, $S_{ч,од,зн,1}$ у ланцюжках 4–10) і об'єднаємо однотипні кроки 4–7 (рис. 4).

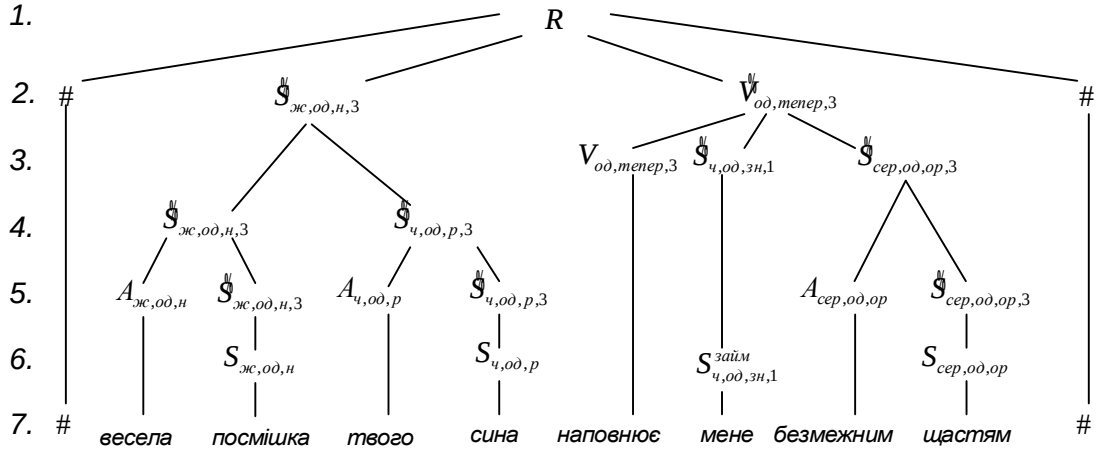


Рис. 4. Приклад контекстозалежної граматики безпосередніх складових (тип 1)

Контекстовільні граматики є окремим випадком граматик безпосередніх складових (прикладом використання контексту в G_1 є правило II.3). Їх цінність зумовлена такими обставинами:

- відмова від контексту (в лівій частині правила є рівно один символ) робить структуру граматик простішою та полегшує їх дослідження;
- хоча в природних мовах заміна одних одиниць іншими допустима в певних контекстах, доцільно досліджувати можливість опису мови, відволікаючись від вказаного факту.

Це розмежовує випадки з необхідністю використання контексту, і випадки, де можна обійтися без нього. Особливий інтерес викликає дослідження ситуацій, де контекст змістово необхідний, але формально його враховують за допомогою контекстовільних правил, тобто не розглядають як контекст (вводять в граматику нові категорії). Так, в граматиці G_1 контекстозалежне правило II.3 усувають (алг. 3).

Алгоритм 3. Перетворення контекстозалежної граматики на контекстовільну граматику.

Крок 1. У допоміжний словник вводять нові символи $\mathbb{S}'_{x,y,z}$ для інтерпретації незайменникових іменних груп, на відміну від символів $\mathbb{S}_{x,y,z}$, які позначають довільні іменні групи.

Крок 2. Правило II.3 замінюють двома новими правилами: $\mathbb{S}_{x,y,z,w} \to S^{займ}_{x,y,w}$ і $\mathbb{S}_{x,y,z,3} \to \mathbb{S}'_{x,y,z}$.

Крок 3. У правилах II.1, II.2 і II.4 всі входження символів $\mathbb{S}_{x,y,z,3}$ замінюють символами $\mathbb{S}'_{x,y,z,w}$. У разі розгортання довільної іменної групи $\mathbb{S}_{x,y,z,w}$ в конструкцію $A+S$ або $S+S_p$ стежать, щоб в заголовку конструкції не з'явився особовий займенник типу *я*, *ви*, *ти*, *він*, *вона*, *воно*, який не може мати при собі визначень (A або S_p: *новий я* або *ми посмішки*). Це визначають різними способами.

1. Особові займенники вважають іменниками особового класу – $S^{займ}$, розглядають їх як іменні групи $\left(\mathbb{S}\right)$ поряд зі звичайними іменниками. Переходити від іменної групи $\mathbb{S}$ до $S^{займ}$ дозволено лише за умови, що це $\mathbb{S}$ раніше не відокремило від себе A вліво або S_p вправо (правила II.1 та II.2), тобто якщо зліва від символу $\mathbb{S}$ немає прикметника, а справа немає групи іменника в родовому відмінку. Ця умова врахована в правилі II.3.

2. Займенники вважають особливим класом іменників, але поряд з категорією довільна іменна група $\mathbb{S}$ вводять категорію власне іменної (незайменникової) групи $\mathbb{S}'$, символ $\mathbb{S}$ під час виведення – до його розгортання – обов'язково замінюють або на символ $S^{займ}$ (який не може розгортатися далі), або на символ $\mathbb{S}'$ (який розгортається звичайно); A та S_p з'являються лише із $\mathbb{S}'$, але $\mathbb{S}'$ не може перетворитися на займенник.

3. Займенники не вважають іменниками та використовують для них символ M. Тоді більшість правил граматики G_1 дублюють, наприклад, поряд з правилом I вводять правило I': $R \to M_{од,наз,w}\mathbb{V}_{од,тепер,w}$; поряд з правилом III.3 – правило III.3': $\mathbb{V}_{у,тепер,w} \to V_{у,тепер,w}M_{х',у',зн,w'}$ тощо. Отримана граматика буде контекстовільною.

Розглянутий приклад показує, що в природних мовах можливі ситуації, коли явища, залежні від контексту, описують і як незалежні від контексту, тобто в термінах контекстовільних граматик. При цьому опис ускладнюють, вводячи нові категорії та правила. Не кожну контекстозалежну замінюють еквівалентною контекстовільною граматикою. Відомо, що існують мови безпосередніх складових, що не є контекстовільними мовами, наприклад, мови вигляду $a^n b^n a^n$ $\left(aba, aabbaa, \mathbf{K}\right)$ або $a^n b^n c^n$. Майже всі приклади мов безпосередніх складових, що не є контекстовільними мовами, мають абстрактний характер і не мають інтерпретацій в природних мовах.

Граматика G_1 не дає бінарної структури. Це відбувається завдяки правилам III.1 і III.2, які відображають таке інтуїтивне розуміння будови речення, за якого група присудка є такою, що складається з особового дієслова і груп його доповнень.

Однак для будь-якої контекстовільної граматики можна побудувати еквівалентну їй бінарну контекстовільну граматику. Наприклад, контекстовільну граматику, подану на рис. 2, перетворюють на бінарну, замінюючи правила III.1 і III.2 такими новими правилами:

III. 1') $V_{y,тепер,w} \to V^1_{y,тепер,w} S_{x'',y'',op,w'''}$; III. 1") $V^1_{y,тепер,w} \to V_{y,тепер,w} S_{x',y',зн,w'}$;

III. 2') $V_{y,тепер,w} \to V^2_{y,тепер,w} S_{x'',y'',зн,w'''}$; III. 2") $V^2_{y,тепер,w} \to V_{y,тепер,w} S_{x',y',op,w'}$.

Необхідно ще замінити правило I правилом I': $R \to S_{x,y,н,w} V_{y,тепер,w}$; тим самим усуваємо межові символи (взагалі в контекстовільній граматиці межові символи формально не потрібні, тоді як в граматиці безпосередніх складових, що має контекстозалежні правила, межові символи можуть бути необхідні як контекст (правило II.3 в G_1). Введене обмеження (не більше від двох символів у правій частині правил) можна накласти і на довільну граматику безпосередніх складових, формулюючи його дещо інакше: кожне правило має вигляд $Z_1 C Z_2 \to Z_1 W Z_2$, де W складається з одного або двох символів. Таку граматику безпосередніх складових природно також назвати бінарною. Для будь-якої граматики безпосередніх складових можна побудувати еквівалентну їй бінарну граматику безпосередніх складових (рис. 5).

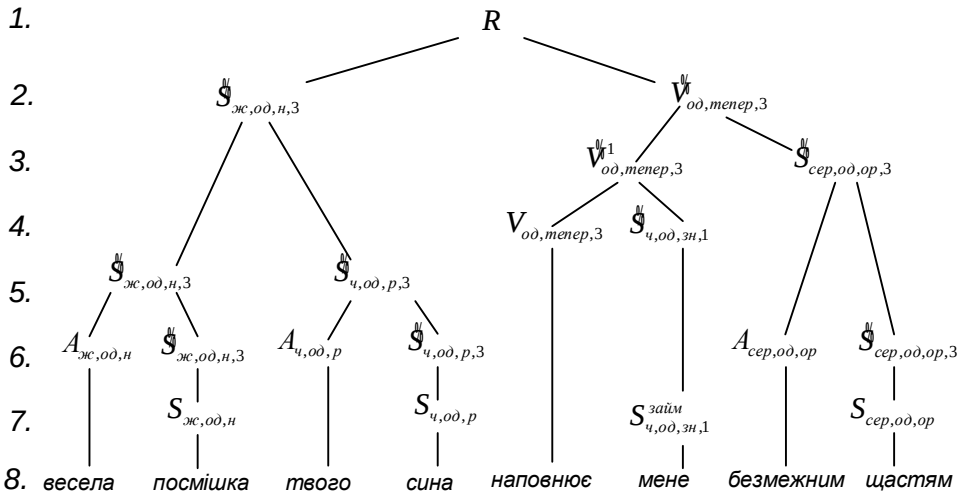


Рис. 5. Приклад контекстовільної граматики (тип 2)

Особливістю регулярних граматик є специфічна форма виведення. Побудуємо для прикладу регулярну граматику G_2, тобто породження речень типу *Весела посмішка наповнює безмежним щастям* (спрощений варіант речення з рис. 1).

Схема граматики G_2.

1) $R \to S_{x,y,н,w}$
2) $S_{x,y,z} \to весела_{x,y,z} S_{x,y,z}$
3) $S_{x,y,z} \to безмежний_{x,y,z} S_{x,y,z}$
4) $S_{ж,y,н} \to посмішка_{ж,y,н} V_{y,3}$
5) $S_{сер,y,op} \to щастя_{сер,y,op} V_{y,3}$
6) $S_{ж,y,н} \to посмішка_{ж,y,н}$
7) $S_{сер,y,op} \to щастя_{сер,y,op}$
8) $V_{y,3} \to наповнити_{y,3} S_{x,y',op}$

Вказане речення матиме в цій граматиці таке виведення:є

R

(1) $S_{ж,од,н}$

(2) *весела* $S_{ж,од,н}$

(4) *весела посмішка* $V_{од,3}$

(8) *весела посмішка наповнює* $S_{сер,од,зн}$

(3) *весела посмішка наповнює безмежним* $S_{сер,од,зн}$

(7) *весела посмішка наповнює безмежним щастям*.

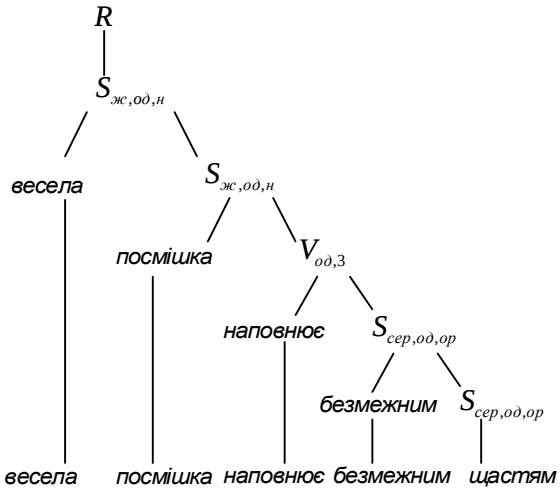


Рис. 6. Приклад регулярної граматики (тип 3)

Кожен проміжний ланцюжок містить рівно один допоміжний символ на останньому місці. Речення породжується зліва направо: на кожному кроці видається конкретна словоформа, а за нею – допоміжний символ, що вказує, яка конструкція повинна стояти за цією словоформою. Потім видається словоформа, що починає цю конструкцію або міститься в ній, після чого знову слідує допоміжний символ чергової конструкції тощо. Регулярна граматика передбачає, що слідує за виданою словоформою, причому глибина передбачення – один сусідній символ; кожен черговий вибір повністю обумовлений одним попереднім вибором. Зазначимо, що із виведення речення в регулярній граматиці неможливо отримати природне подання структури безпосередніх складових цього речення (як це робилося для контекстозалежної та контекстовільної граматик). Тобто регулярні граматики дають деяку структуру складових, однак ці сладові зазвичай є суто формальними і не піддаються природній інтерпретації (рис. 6).

Некоректне розбиття речення на дві складові – *весела* тощо, а також з приписуванням категорій отриманим складовим. У реченні *Посмішка наповнює мене щастям* результат був би ще гірший: складовою є поєднання *мене щастям*. Це речення не породжується граматикою G_2, проте неважко доповнити її (введенням двох правил). Інтерпретація виведення в регулярних граматиках не має сенсу. Використовують іншу інтерпретацію регулярного виведення: як послідовності передбачень та їх реалізацій. Існують контекстовільні мови, що не породжуються регулярними граматиками. Прикладом слугує мова, що складається з ланцюжків вигляду $a^n b^n$.

Необмежені граматики типу 0 є лише окремим випадком загального поняття граматики. Проте вони, безумовно, достатні для опису будь-яких природних мов у повному обсязі. Будь-яка природна мова (множина правильних фраз) є легкорозпізнаваною множиною. Це означає існування доволі простого алгоритму розпізнавання правильності фраз. Якщо ж мова розпізнається алгоритмом з вказаним обмеженням на об'єм 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 перевершує за довжиною числа Kn (K – деяка константа). Така граматика є *граматикою з обмеженим розтягуванням*, де ємнісна сигнальна функція не більша від лінійної. 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 контекстовільних граматик, легко описуються граматикою G_0.

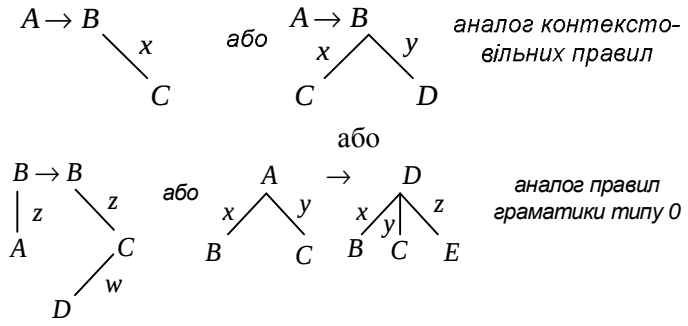
Недоліки *методу виведення* граматики G_0 зводяться до трьох пунктів.

1. За їх допомогою неможливо 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 точно певним порядком слів – із тим, який 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, заданого самим деревом, є ще і відношення лінійного порядку (правіше – лівіше). Тобто в синтаксичній структурі граматики G_0 не розчленовані 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. Порядок слів залежить від синтаксичної структури; він визначається обов'язково з її урахуванням, тому є щодо неї чимось похідним, вторинним. Доцільно видозмінити поняття граматики, що породжує, так, щоб ліві й праві частини правил підстановки були не лінійно впорядкованими ланцюжками, а наприклад, дерева (без лінійної впорядкованості), що відображають синтаксичні відношення [18–20]. Тоді правила мають такий вигляд:



Риски з індексами зображають синтаксичні зв'язки різних типів; літери $A,B,C,...$ – синтаксичні категорії. NB: взаємне розташування символів одного рівня підпорядкування не відіграє жодної ролі і є на цій схемі випадковим 

; означає те саме, що і 

[18–20].

У результаті отримують обчислення синтаксичних структур (а не фраз) мови. Це обчислення є частиною граматики, що породжує [19]. Іншу частину цієї граматики становить обчислення, яке для будь-якої цієї синтаксичної структури задає (з урахуванням яких-небудь інших факторів, наприклад, в українській мові – з обов'язковим урахуванням логічного виділення тощо) всі можливі для неї лінійні послідовності слів. Тоді знімається проблема розривних складових [19]. 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.

Під час аналізу досліджують багаторівневу структуру текстового контенту: лінійну послідовність символів; лінійну послідовність морфологічних структур; лінійну послідовність речень; мережу взаємопов'язаних єдностей (алг. 4).

Алгоритм 4. Лінгвістичний аналіз текстового комерційного контенту.

Етап. 1. Граматичний аналіз текстового контенту.

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

Крок 2. Поділ ланцюжка символів на слова.

Крок 3. Виділення цифр, чисел, дат, незмінних зворотів і скорочень.

Крок 4. Видалення нетекстових символів.

Крок 5. Формування та аналіз лінійної послідовності слів зі службовими знаками (алг. 6).

Етап. 2. Морфологічний аналіз текстового контенту.

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

Крок 2. Кожній словоформі ставиться у відповідність значення граматичних категорій (сукупності граматичних значень: рід, відмінок, відмінювання тощо).

Крок 3. Формування лінійної послідовності морфологічних структур.

Етап. 3. Синтаксичний аналіз текстового контенту (алг. 5).

Етап. 4. Семантичний аналіз текстового контенту.

Крок 1. Слова співвідносяться з семантичними класами зі словника.

Крок 2. Відбір потрібних для цього 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 контенту. 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. 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.

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, побічно пов'язані з внутрішньою структурою, тому не розглядаються.

За допомогою бази даних (бази термінів/морфем і службових частин мови) та визначених правил аналізу тексту виконують пошук терміна. Синтаксичні аналізатори працюють в два етапи: ідентифікують змістові лексеми та створюють дерево розбору (алг. 5).

Алгоритм 5. Синтаксичний аналізатор комерційного текстового контенту.

Етап. 1. Ідентифікація змістових лексем.

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. Кожен крок виведення полягає або в розгортанні одного з символів попереднього ланцюжка, або в заміні його іншим, інші ж символи переписуються без зміни. Якщо під час розгортання, замінювання або переписування символи (*предки*) з'єднаємо лініями безпосередньо з символами, які утворюються в результаті розгортання, заміни або переписування (*нащадками*), отримаємо дерево складових, або синтаксичну структуру.

Крок 1. Розгортання іменної групи. Розгортання дієслівної групи.

Крок 2. Реалізація синтаксичних категорій словоформами.

Етап. 3. Визначення множини ключових слів.

Крок 1. Визначення термів *Noun* – іменників, словосполучень іменників або прикметника з іменником серед множини слів текстового комерційного контенту.

Крок 2. Розрахунок унікальності *Unicity* для термів *Noun*.

Крок 3. Розрахунок *NumbSymb* (кількість знаків без пробілів) для *Noun* при $Unicity \geq 80$.

Крок 4. Розрахунок *UseFrequency* – частоти вживання ключових слів. Для термів з $NumbSymb \leq 2000$ частота *UseFrequency* в межах $(6;8]$ %., з $NumbSymb \geq 3000$ частота *UseFrequency* в межах $[2;4)$ %, з $2000 > NumbSymb < 3000$ частота *UseFrequency* є $[4;6]$ %.

Крок 5. Розрахунок *BUseFrequency* – частота вживання ключових слів на початку тексту, *IUseFrequency* – частота вживання ключових слів у середині тексту, *EUseFrequency* – частота вживання ключових слів у кінці тексту.

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

Крок 7. Сортування ключових слів згідно з їхніми пріоритетами.

Етап. 4. Заповнення бази пошукових образів контенту, тобто атрибутів *KeyWords* – 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 у середині тексту, *EUseFrequency* – частота вживання ключових слів у кінці тексту.

Керуючись правилами породжувальної граматики, виконують корекцію терміна згідно з правилами його вживання у контексті. Речення задають межі дії знаків пунктуації, анафоричних і катафоричних посилань. Семантика тексту зумовлена комунікативним завданням передавання інформації. Структура тексту визначається внутрішньою організацією одиниць тексту і закономірностями їх взаємозв'язку. Під час синтаксичного аналізу текст оформляють у структуру даних, наприклад, у дерево, яке відповідає синтаксичній структурі вхідної послідовності, і найкраще підходить для подальшого опрацювання. Після аналізу фрагмента тексту і терміна синтезують новий термін як ключове слово тематики контенту, використовуючи базу термінів та їх морфем. Далі синтезують терміни для формування нового ключового слова, використовуючи базу службових частин мови. Принцип виявлення ключових слів за змістом (термами) ґрунтується на законі Зіпфа і зводиться до вибору слів із середньою частотою появи (найвживаніші слова ігнорують через "стоп-словники", а рідкісні слова тексту не враховують). За змістовий аналіз контенту відповідає процес видобування граматичних даних зі слова через графемний аналіз та корегування результатів морфологічного аналізу через аналіз граматичного контексту лінгвістичних одиниць (алг. 6).

Алгоритм 6. Рубрикація текстового комерційного контенту

Етап. 1. Поділ комерційного контенту на блоки.

Крок 1. Подання на вхід блока побудови дерева блока комерційного контенту.

Крок 2. Створення нового блока 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 блока та перехід до крок 3.

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

Крок 6. Отримання на виході дерева блоків у вигляді таблиці.

Етап. 2. Поділ блока на речення зі збереженням структури.

Крок 1. На вхід подається таблиця блоків. Створення таблиці речень зі зв'язком за полем Код_розділу типу *n-to-*1 із таблицею блоків.

Крок 2. Створення нового речення в таблиці речень.

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, розбір нового речення та перехід до кроку 2.

Крок 5. Перевірка наявності кінця тексту блока. Якщо кінець тексту, то перехід до кроку 6, якщо ні, то збереження послідовності у таблиці, розбір нового речення та перехід до кроку 2.

Крок 6. Отримують на виході дерево речень у вигляді таблиці.

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

Крок 8. Отримання на виході дерева речень у вигляді таблиць.

Етап. 3. Поділ речень на лексеми із вказанням належності до речень.

Крок 1. Формування на основі таблиці речень таблиці лексем із полями Код_лексеми (унікальний ідентифікатор), Код_речення (число, що дорівнює коду речення з лексемою), Номер_лексеми (число, що дорівнює номеру лексеми в реченні), Текст (текст лексеми).

Крок 2. Подання на вхід для розбору на лексеми речення з таблиці речень.

Крок 3. Створення нової лексеми в таблиці лексем.

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, розбір нової лексеми та перехід до кроку 3.

Крок 6. Проведення синтаксичного аналізу на основі даних, одержаних на виході (алг. 5).

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

Етап. 4. Визначення тематики комерційного контенту.

Крок 1. Побудова ієрархічної структури властивостей 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 (алг.5).

Крок 5. Визначення *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 вживання 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.

Крок 9. Присвоєння визначеній рубриці аналізованого комерційного контенту.

Крок 10. Розрахунок *Location* – коефіцієнт розташування контенту в 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 текстового контенту, *Static* – коефіцієнт статистичної важливості термів, *Addterm* – коефіцієнт наявності додаткових термів, *TKeyWords* – тематичні ключові слова, *FKeyWords* – частота вживання ключових слів, *Comparison* – порівняння вживання ключових слів різних тематик, *QuantitativeryTKey* – частота вживання тематичних ключових слів у тексті комерційного контенту.

Побудова тексту визначається темою, вираженою інформацією, умовами спілкування, завданням повідомлення та стилем викладення. Із семантичною, граматичною та композиційною структурою контенту пов’язані його стильові/стилістичні характеристики, залежні від індивідуальності автора та підпорядковані тематичній/стильовій домінанті тексту.

Основні етапи визначення морфологічних ознак одиниць тексту: 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 через автоматичне індексування складових комерційного контенту поділений на послідовні блоки: морфологічний аналіз, синтаксичний аналіз, семантико-синтаксичний аналіз лінгвістичних конструкцій та варіювання змістового запису текстового контенту.

Використано такі способи вираження граматичного значення: синтетичний, аналітичний, аналітико-синтетичний та суплетивний. Граматичні значення узагальнені 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'єднувані у парадигми для класифікації частин тексту. Об'єктом морфологічного аналізу є структура слова, форми словозміни, способи вираження 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'язане з граматичною класифікацією. Морфорлогічний 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 забезпечення пошуку для певного поняття загального/асоційованого з ним поняття; система обслуговування словників. Процес індексування залежить від дескрипторного словника або інформаційно-пошукового тезаурусу. Дескрипторний словник має структуру таблиці з трьома колонками: основи слів; набори дескрипторів, приписані кожній основі; граматичні ознаки дескрипторів. Індексування складається з: виділення інформативних словосполучень з тексту; розшифрування абревіатури; заміни слів з основами-дескрипторами на код дескриптора; зняття омонімії.

Висновки

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, що запропонував 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. В роботі розглянуто вплив норм та правил мови на хід 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, описані як незалежні від контексту, тобто в термінах контекстовільних граматик. При цьому опис ускладнений через утворення нових категорій і правил. В статті визначено особливості процесу введення нових обмежень на класи цих граматик через введення нових правил. За кількості символів у правій частині правил, не меншій за ліву, отримали нескорочені граматики. Потім заміною лише одного символу одержано контекстозалежні граматики. За наявності в лівій частині правила лише одного символу отримали контекстовільні граматики. Жодних інших 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 формувати системи синтезу мови та текстів, а також створювати підручники практичної морфології, таблиці словозміни, укладати списки морфем (афіксів, коренів), визначати продуктивності та частотності морфем, встановлювати 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 інформаційних пошукових системах тощо. Корисно вводити все нові й нові обмеження на ці граматики, отримуючи вужчі їх класи.

Описуючи складне коло явищ, обмежують набір використовуваних засобів опису, розглядаючи і такі засоби, які подають в загальному випадку свідомо недостатніми. Дослідження починають із мінімальних засобів; щоразу, коли їх недостатньо, поступово вводять (дрібнішими порціями) нові засоби, завдяки чому вдається точно визначити, якими засобами можна/не можна обійтися в описі того або іншого явища для розуміння його природи.

У статті розглянуто відомі способи і підходи до вирішення проблеми автоматичного опрацювання текстового контенту та виділено недоліки й переваги різних підходів та результатів у галузі синтаксичних аспектів комп'ютерної лінгвістики. Сформовано загальні концептуальні принципи моделювання словозмінних процесів під час утворення 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. Застосування методики дає змогу досягти вищих показників надійності порівняно з відомими аналогами, а також демонструє високу ефективність у прикладних застосуваннях для побудови нових інформаційних технологій лексокографування та дослідження словозмінних ефектів природних мов. Робота має практичну цінність, оскільки запропоновані моделі та правила дають можливість ефективно організовувати 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.: 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. Непосредственных 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. Багмут А.Й. Порядок слів / А.Й. Багмут // Українська мова: енцикл. – 3-тє вид., зі змінами і доп. – К.: Укр. енциклопедія, 2007. – С. 675–676. 7. Берко А. Системи електронної контент-комерції / А. Берко, В. Висоцька, В. Пасічник. – Л: НУЛП, 2009. – 612 с. 8. Берко А.Ю. Застосування методу контент-аналізу для формування інформаційних ресурсів в системах електронної контент-комерції / А.Ю. Берко, В.А. Висоцька, М.М. Сороковський // "Інформаційні системи та мережі". Вісник Національного університету "Львівська політехніка". – Львів, 2012. – № 743. – С. 3–15. 9. Бильгаева Н.Ц. Теория алгоритмов, формальных языков, грамматик и автоматов: учебное пособие / Н.Ц. Бильгаева. – Улан-Удэ: Изд-во ВСГТУ, 2000. – 51 с. 10. 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. 11. Брайчевский С. Современные информационные потоки / С. Брайчевский, Д. Ландэ // Научно-техническая информация. – 2005. – № 11. – С. 21–33. 12. Висоцька В.А. Застосування породжувальних граматик для моделювання синтаксису речення / В.А. Висоцька, Т.В. Шестакевич, Ю.М. Щербина // "Інформаційні системи та мережі". Вісник Національного університету "Львівська політехніка". – Львів, 2012. – № 743. – С. 175–190. 13. Висоцька В.А. Утворення речень англійською та німецькою за допомогою породжувальних граматик / В.А. Висоцька, Т.В. Шестакевич, Ю.М. Щербина // "Комп'ютерні науки та інформаційні технології". Вісник Національного університету "Львівська політехніка". – Львів, 2012. – № 744. – С.142–152. 14. Висоцька В.А. Генерування речень українською за допомогою породжувальних граматик / В.А. Висоцька, Т.В. Шестакевич // Міжнародна наукова конференція "Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту (ISDMIT'2012)", Євпаторія. – 27–31 травня 2012. – С. 48–50. 15. Волкова И.А. Формальные грамматики и языки. Элементы теории трансляции: учеб. пособ. для студентов II курса / И.А. Волкова, Т.В. Руденко. – 2-е изд., перераб. и доп. – М.: Издательский отдел факультета вычислительной математики и кибернетики МГУ им. М.В.Ломоносова, 1999. – 62 с. 16. 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. 17. Герасимов А.С. Лекции по теории формальных языков / А.С. Герасимов. – Режим доступа: http://gas-teach.narod.ru/au/tfl/tfl01.pdf. 18. Гладкий А.В. Синтаксические структуры естественного языка в автоматизированных системах общения / А.В. Гладкий. – М.: Наука, 1985. – 144 с. 19. Гладкий А.В. Элементы математической лингвистики / А.В. Гладкий, И.А. Мельчук. – М.: Наука, 1969. – 192 с. 20. Гладкий А.В. Формальные грамматики и языки / А.В. Гладкий. – М.: Наука, 1973. – 368 с. 21. Гросс М. Теория формальных грамматик / М. Гросс, А. Лантен; пер. с франц. И.А. Мельчука под ред. А.В. Гладкого. – М.: Мир, 1971. – 294 с. 22. Дарчук Н.П. Комп'ютерна лінгвістика (автоматичне опрацювання тексту): підручник / Н.П. Дарчук. – К.: ВПЦ "Київський університет", 2008. – 351 с. 23. 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. 24. Зубков М. Українська мова: універсальний довідник / М. Зубков. – К.: ВД "Школа", 2004. – 496 с. 25. Корнеев В. Базы данных. Интеллектуальная обработка информации / В. Корнеев, А. Гареев, С. Васютин, В. Райх. – М: Нолидж, 2000. – 352 с. 26. Ландэ Д. Основы моделирования и оценки электронных информационных потоков / Д. Ландэ, В. Фурашев, С. Брайчевский, О. Григорьев. – К: Інжиніринг, 2006. – 348 с. 27. Ландэ Д. Основы интеграции информационных потоков: монография / Д. Ландэ. – К: Інжиніринг, 2006. – 240 с. 28. 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. – Київ, 2011. – 19 с. 29. Мартыненко Б.К. Языки и трансляции: учеб. пособие / Б.К. Мартыненко. – 2-е изд., испр. и доп. – СПб.: Изд-во С.-Петерб. ун-та, 2008. – 257 с. 30. 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. 31. Носков С.А. Самоучитель немецкого языка. Deutsch für sie / С.А. Носков. – К.: Наука, 1999. – 400 с. 32. Партико З.В. Прикладна і комп'ютерна лінгвістика. Вступ до спеціальності: навч. посіб. / З.В. Партико. – Л.: Афіша, 2008. – 224 с. 33. Пасічник В.В. Математична лінгвістика. Книга 1. Квантитативна лінгвістика: навч. посіб. / В.А. Висоцька, В.В. Пасічник, Ю.М. Щербина, Т.В. Шестакевич. – Львів: Новий світ-2000, 2012. – 359 с. 34. Пентус А.Е. Теория формальных языков: учебное пособие / А.Е. Пентус, М.Р. Пентус. – М.: Изд-во ЦПИ при механико-математическом ф-те МГУ, 2004. – 80 с. 35. Попов Э.В. Общение с ЭВМ на естественном языке / Э.В. Попов. – М.: Наука, 1982. – 360 с. 36. Постнікова О.М. Німецька мова. Розмовні теми: лексика, тексти, діалоги, вправи / О.М. Постнікова. – К.: А.С.К, 2001. – Т. 1. – 400 с. 37. Постнікова О.М. Німецька мова. Розмовні теми: лексика, тексти, діалоги, вправи / О.М. Постнікова. – К.: А.С.К, 2001. – Т. 2. – 320 с. 38. 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. 39. Русаченко Н.П. Морфонологічні процеси у словозміні та словотворі староукраїнської мови другої половини XVI – XVIII ст.: автореф. дис. на здобуття наук. ступ. канд. філол. наук: спец. 10.02.01 / Н.П. Русаченко. – К., 2004. – 24 с. – Режим доступу: http://auteur.corneille-moliere.com/?p=history&m=corneille_moliere&l=rus. 40. 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. 41. Туришева О.О. Порушення рамкової конструкції в сучасній німецькій мові: функціональний аспект, нормативний статус: автореф. дис. канд. філол. наук: спец. 10.02.04 / Туришева О.О. – Одеса, 2012. – 20 с. 42. Український правопис / Ін-т мовознавства ім. О.О. Потебні НАН України, Ін-т укр. мови НАН України. – К.: Наук. думка, 2007. – 288 с. 43. Федорчук А. Г. Контент-мониторинг информационных потоков [Электронный ресурс] / Б-ка Нац. акад. наук: пробл. функционирования, тенденции развития. — К., 2005. — Вып. 3. — Режим доступа: http://www.nbuv.gov.ua/articles/2005/05fagmip.html. — Загол. с экрана. 44. Фомичев В.С. Формальные языки, грамматики и автоматы / В.С. Фомичев. – Режим доступа: http://www.proklondike.com/books/thproch/. 45. Хомский Н. О некоторых формальных свойствах грамматик / Н. Хомский // Кибернетический сборник. – М.: Мир, 1962. – № 5. – С. 279–311. 46. Хомский Н. Формальный анализ естественных

языков / Н. Хомский, Дж. Миллер // Кибернетический сборник. – М.: Мир, 1965. – № 1. – С. 231–290. 47. Хомский Н. Язык и мышление / Н. Хомский // Публикации ОСиПЛ. Серия монографий. – М.: Издательство Московского университета, 1972. – № 2. – 122 с. 48. Хомский Н. Синтаксические структуры / Н. Хомский // Сборник "Новое в лингвистике". – М.: ИЛ, 1962. – № 2. – С. 412–527. 49. 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. 50. Шаров С.А. Средства компьютерного представления лингвистической информации / С.А. Шаров. – Режим доступу : http://www.ksu.ru/eng/science/ittc/vol000/002/. 51. 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. 52. Шульжук К. Синтаксис української мови: Підручник / К. Шульжук. – К.: Академія, 2004. – 397 с. 53. Щербина Ю.М. Предмет математичної лінгвістики / Ю.М. Щербина // Вісник НУЛП "Інформаційні системи та мережі". – Львів, 2002. – № 464. – С. 340–349. 54. Щербина Ю.М. Науковий напрям та навчальна дисципліна "Математична лінгвістика" / Ю.М. Щербина, Т.В. Шестакевич, В.А. Висоцька // Вісник НУЛП "Інформаційні системи та мережі". – Львів, 2010. – № 673. – С. 384–392. 55. Щербина Ю.М. Утворення українських дієприкметників за допомогою породжувальних граматик / Ю.М. Щербина, Ю.В. Нікольський, В.А. Висоцька, Т.В. Шестакевич // "Інформаційні системи та мережі". Вісник Національного університету "Львівська політехніка". – 2011. – № 715. – С. 354–369. 56. 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 языка / Н. Хомский // Кибернетический сборник. – М.: ИЛ, 1961. – № 2. – С. 237–266). 57. 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). 58. Chomsky N. On the notion "Rule of Grammar" / N. Chomsky // Proc. Symp. Applied Math., 12. Amer. Math. Soc., 1961. (Русский перевод: Хомский Н. О понятии "правило грамматики" / Н. Хомский // Сб. Новое в лингвистике. – М.: Прогресс, 1965. – № 4. – С. 34–65). 59. Chomsky N. Context-free grammars and pushdown storage / N. Chomsky // Quarterly Progress Reports, № 65, Research Laboratory of Electronics, M.I.T., 1962. 60. Chomsky N. Formal properties of grammars / N. Chomsky // Handbook of Mathemati-Mathematical Psychology, 2, ch. 12, Wiley, 1963. – P. 323–418. (Русский перевод: Хомский Н. Формальные свойства грамматик / Н. Хомский // Кибернетический сборник. – М.: ИЛ, 1966. – № 2. – С. 121–230). 61. 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). 62. Chomsky N. Finite state languages / N. Chomsky, G.A. Miller // Information 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). 63. 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). 64. 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). 65. English Verbs (Part 1) – Basic Terms. – Режим доступу: http://sites.google.com/site/englishgrammarguide/Home/ english-verbs--part-1----basic-terms. 66. Vysotska Victoria. Web Content Processing Method for Electronic Business Systems / Victoria Vysotska, Lyubomyr Chyrun // International Journal of Computers & Technology – PP. 3211–3220. ISSN 2277-3061, Impacr factor 1,341. http://cirworld.com/index.php/ijct/article/view/3299. (Index Copernicus, NASA ADS, DOAJ, Google Scholar, Eyesource, EBSCO, CiteSeer, UlrichWeb, ScientificCommons).