УДК 81:004.93

Т. В. Шестакевич, В. А. Висоцька, Л. В. Чирун, Л. Б. Чирун
Національний університет “Львівська політехніка”,
кафедра інформаційних систем та мереж

МОДЕЛЮВАННЯ СЕМАНТИКИ РЕЧЕННЯ ПРИРОДНОЮ МОВОЮ ЗА ДОПОМОГОЮ ПОРОДЖУВАЛЬНИХ ГРАМАТИК

Розглянуто 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

Активний розвиток Інтернету сприяє створенню різноманітних лінгвістичних ресурсів. Необхідність у реалізації процесів аналізу та синтезу природномовних текстів зумовила появу відповідних лінгвістичних моделей процесів їх опрацювання, оскільки розвиток багатьох мовознавчих дисциплін потрібен для інформаційних наук. Інтеграційні процеси в більшості галузей життя сучасного світу привертають особливу увагу до розроблення та створення автоматизованих систем опрацювання багатомовної інформації.

Аналіз останніх досліджень і публікацій

Лінгвістичний аналіз природномовних текстів складається з декількох послідовних процесів: графемного, морфологічного, синтаксичного та семантичного аналізу. Для кожного з цих етапів створено відповідні моделі та алгоритми. Теорія породжувальних граматик, початок якої закладено у роботах американського лінгвіста Н. Хомські [12–16, 38–41, 49–57], є ефективним інструментом лінгвістичного моделювання на синтаксичному рівні мови. Н. Хомські використав прийом формального аналізу граматичної структури фраз, який дає змогу виділити синтаксичні структури (складові), що є основною схемою фрази, незалежно від її значення [12]. Ідеї Н. Хомські розвивав, серед інших, радянський лінгвіст А. В. Гладкий [14–16], який застосовував поняття дерев залежності та систем складових для моделювання синтаксичного рівня мови [22–24, 43–48]. Він запропонував спосіб моделювання синтаксису за допомогою синтаксичних груп, що виділяють складові словосполучень як одиниці побудови дерева залежностей, – таке подання дало змогу об’єднати переваги методу безпосередніх складових і дерев залежностей [4, 5, 43–48]. Напрацювання Н. Хомські [35–38, 49–57] та А. В. Гладкого [14–16], дослідження М. Гросса і А. Лантена [17], А. В. Анісімова [2–3], Ю.Д. Апресяна [4–5], Н. Ц. Більгаєвої [8], І.А. Волкової та Т. В. Руденка [11], Є. І. Большакової, Е. С. Клишинського, Д.В. Ланде, А. А. Носкова, О. В. Пескової та Є. В. Ягунової [9], А. С. Герасимова [13], Б. К. Мартиненко [23], А. Є. Пентуса та М. Р. Пентуса [28], Е. В. Попова [29], В. С. Фомічева [37] застосовні до 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, 22–24, 27, 43, 61, 67].

Формулювання мети

Покажемо способи застосування апарату породжувальних граматик до моделювання синтаксису речень для різних мов: англійської, німецької та української [1–67]. Для цього розберемо синтаксичну структуру речень, продемонструємо особливості процесу синтезу речень зазначених мов. Розглянемо вплив норм та правил мови на процес побудови граматик [10, 21, 26, 44, 46, 47].

Аналіз отриманих наукових результатів

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]. Граматики класифікують за типами продукцій, на які накладено певні обмеження (табл. 1) [38–41].

Таблиця 1

Класифікація граматик за типами продукцій

Граматика	Тип	Опис
G_0	Безпосередніх складових	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.

Тлумачитимемо термінальні символи як словоформи (деякої природної мови), нетермінальні символи – як синтаксичні категорії, а термінальні ланцюжки, що виводяться, – як правильні речення даної мови [12–16, 38–41, 49–57]. Тоді виведення речення природно інтерпретується як його синтаксична структура, яка подана в термінах безпосередній складових, тобто способом, давно відомим у лінгвістиці [10, 21, 26, 44, 46, 47]. Пояснимо сказане прикладами [14–16].

Для регулярної граматики G_3 характерні такі властивості [14–16]:

1. Породжує фрази строго в одному напрямі (зліва направо), розгортаючи їх слово за словом.
2. Володіє *короткою пам'яттю*, тобто рівно на один крок.

У фразі часто буває так, що слова b і c далеко віддалені один від одного, але володіють певною семантичною відповідністю та залежністю. В граматиках безпосередніх складових G_0 та контекстно-вільних граматиках G_2 це враховують простим і природним чином: достатньо, щоб слова b і c або їх предки з'являлися разом, на одному кроці виведення як безпосередні нащадки одного і того ж символу. Саме у цей момент їм і приписують інформацію про наявність відповідності. Після цього між ними вставляють скільки завгодно інших символів – інформація про відповідність зберігається. Так, в граматиці G_1 предки підмета і особистого дієслова (символи $\tilde{S}_{x,y,наз,w}$ і $\tilde{V}_{y,тепер,w}$) з'являються одночасно як нащадки символу *Sentence* при вживанні правила $Sentence \to \tilde{S}_{x,y,наз,w}\tilde{V}_{y,тепер,w}$, та їх узгодження в числі і особі (y, w) зберігається до кінця виведення, щоб не було вставлено між ними. Тобто інформація про узгодження слів b і c

пам’ятається при будь-якій відстані між ними. У цьому сенсі можна сказати, що граматики G_2 мають необмежену пам’ять. Що ж стосується граматик G_3, то вони в тому ж самому сенсі мають обмежену пам’ять. Найважливішою особливістю регулярних граматик є специфічна форма виведення. Справа в тому, що граматика G_3 здатна передавати інформацію про відповідність тільки безпосередньо від переднього символу до наступного [14–16].

Побудуємо для прикладу регулярну граматику G_3 (табл. 1), тобто породження речень типу *Весела посмішка наповнює безмежним щастям* [6, 7, 10, 14–16, 19, 20, 32, 33, 36, 45, 48–57].

Таблиця 1

Схема граматики G_3 для прикладу 1

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

Вказане речення матиме в цій граматиці таке виведення [14–16]:

Sentence

(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*.

У правилах 4 і 5 граматики G_3 інформація про число (індекс y) передається від підмета безпосередньо до наступного за ним дієслова. Тому, якщо інформацію про відповідність між b і c доводиться передавати через якісь проміжні символи, то в граматиці G_3 це можна зробити тільки приписавши вказівки про наявність відповідності всім проміжним символам, для яких ці вказівки, по суті, не потрібні. Так, якщо необхідно породжувати за допомогою граматики G_3 українські фрази, де підмет відділяється від особистого дієслова якимись словами, наприклад, іменником у родовому відмінку (*посмішка дитини наповнює щастям*), необхідно ввести спеціальні правила, де дієслово формально буде узгоджуватися безпосередньо з передуючим йому іменником у родовому відмінку. Однак таке узгодження буде виконуватися не на основі власних ознак цього іменника, а на основі штучних *невласних* ознак, що змістовно означають число (і рід) підмета – адже саме з ним і повинне узгоджуватися дієслово [14–16]. Це означає, що замість кожного символу $S_{x,y,род}$ треба буде ввести шість нових символів: $S_{x,y,род \| м,ед}$, $S_{x,y,род \| м,мн}$, …, тобто “іменник роду x, в числі y, в род. від., залежне від підмета чол. роду в числі од.”, “іменник роду x, в числі y, в род. від., залежне від підмета чол. роду у числі мн.” тощо. Аналогічно доведеться чинити і у випадку, коли підмет і присудок розділені прислівниками, наприклад, *весела посмішка дитини* ***стрімко*** *наповнює мене безмежним щастям*: тут треба буде ввести спеціальні категорії прислівників, а саме: 1) прислівники, залежні від дієслова, погодженого з підметом чол. роду в числі од., 2) прислівники, залежні від дієслова, погодженого з підметом чол. роду у числі мн. тощо [14–16].

Кожен проміжний ланцюжок містить рівно один допоміжний символ на останньому місці. Це означає, що речення породжується зліва направо: на кожному кроці видається конкретна

словоформа, а за нею – допоміжний символ, що вказує, яка конструкція повинна слідувати за цією словоформою; потім (на наступному кроці) видається словоформа, що починає цю конструкцію або міститься в ній, після чого знову слідує допоміжний символ чергової конструкції тощо. Регулярна граматика ніби передбачає, що може слідувати за вже виданою словоформою, причому глибина передбачення – один сусідній символ; кожен черговий вибір повністю обумовлюється лише одним попереднім вибором. Зазначимо, що із виведення речення в регулярній граматиці неможливо отримати природне подання структури безпосередніх складових цього речення (як це робилося для контекстно-залежної та контекстно-вільної граматик). Тобто, регулярні граматики дають деяку структуру складових, як і взагалі всі граматики безпосередніх складових, однак ці складові зазвичай мають суто формальний характер і не піддаються природній інтерпретації (рис. 1).

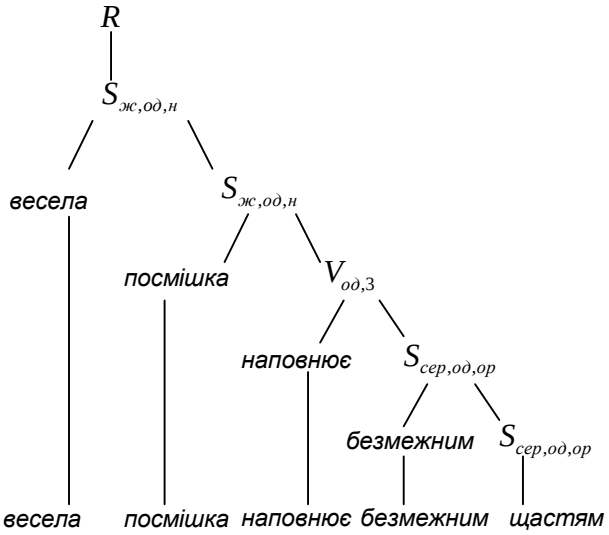


Рис. 1. Приклад 1 регулярної граматики G_3

Навряд чи можна погодитися з розбиттям пропозиції на дві складові – *весела* тощо, а також з приписуванням категорій отриманим складовим. У реченні *Посмішка наповнює мене щастям* результат був би ще гірший: складовою виявилося б поєднання *мене щастям*. Дана пропозиція не породжується граматикою G_2, проте неважко доповнити її так, щоб воно породжувалося (достатньо додати два правила). Тому інтерпретація виведення в регулярних граматиках як структури безпосередніх складових у загальному випадку не має сенсу; зазвичай використовують іншу інтерпретацію регулярного виведення: як послідовності передбачень та їх реалізацій. Нарешті, зазначимо про клас регулярних мов класу контекстно-вільних мов: існують контекстно-вільні мови, що не породжуються регулярними граматиками. Прикладом слугує мова, що складається з ланцюжків вигляду $a^n b^n$. Мови $\{a^n b^n a^n\}$ і $\{a^n b^n c^n\}$ не є контекстно-вільними мовами [14–16].

Більш того, якщо підмет і присудок розділені одним/декількома підрядними реченнями, то кожна з категорій, що зустрічається в цих реченнях, має бути розділена на шість категорій, і, отже, всі правила породження підрядних речень мають бути фактично повторені шість разів. Зрозуміло, все сказане зберігає силу і для тих випадків, коли підмет і присудок розділені багатьма словами, наприклад, скільки завгодно довгим ланцюжком род. відмінків або будь-яким числом підрядних речень. Отже, граматика G_3, як і граматика G_2, здатна забезпечувати узгодження між скільки завгодно далеко віддаленими словами: $\overbrace{a \ldots a'}$. Проте це є громіздким і, головне, дуже неприродним: доводиться запроваджувати багато додаткових категорій (класів слів), таких, що явно суперечать мовній інтуїції [6,7, 10, 14–16, 19, 20, 32, 33, 36, 45, 48–57].

Але це ще не все. Якщо доводиться мати справу не з однією, а з багатьма узгодженими парами, причому кожна наступна “вкладена” в попередню і кількість таких пар теоретично не

обмежене: $abcde...e'd'c'b'a'$, то забезпечити узгодження в подібній ситуації граматика G_3 принципово не здатна. Цей факт може бути строго доведений (див., наприклад, Гладкий, 1966, с. 103, зауваження про мову L_7); викладати саме доведення не будемо, а обмежимося змістовними зауваженнями, що пояснюють, чому це так. Для забезпечення узгодження однієї пари доводиться розділяти всі категорії, що поділяють цю пару; у нашому прикладі з підметом і присудком кількість категорій збільшувалося вшестеро. Якщо ж між словами, що узгоджують, є ще одна пара $ab...b'a'$, то кожну, що узгоджується, з вже поділених проміжних категорій треба буде поділити ще раз, що знову збільшить кількість категорій (кожна категорія, що зустрічається між b і b', повинна буде нести вказівки про узгодження як a із a', так і b із b'). Отже, якщо кількість вкладених пар потенційно нічим не обмежена, то для узгодження потрібно було б мати нескінченно багато категорій (допоміжних символів) – тоді як кількість символів у будь-якій граматиці скінченна. Отже, вказана ситуація з граматикою G_3 описуватися не може [14–16]. Тим часом ця ситуація достатньо типова для природних мов. Так, вона має місце в складних реченнях [6, 7, 10, 14–16, 19, 20, 32, 33, 36, 45, 48–57], наприклад: *Дитина^{a}, якій мама^{b}, що зібралася^{c}, коли тато^{d}, який знаходився^{e}, …, в сусідній кімнаті$^{e'}$, спав$^{d'}$, ідти на роботу$^{c'}$, взула чоботи^{b}, побігла швидко в школу^{a}.* На місце "…" можна вставити будь-які речення, зокрема такі, що містять багато додаткових речень, також із послідовним вкладеннями. Подібні речення регулярно зустрічаються в різних мовах. Приклади потенційно необмеженого вкладення пар, що узгоджуються, можна вказати і для простих речень [6, 7, 10, 14–16, 19, 20, 32, 33, 36, 45, 48–57], зокрема:

1. Слов'янські конструкції з послідовним вкладеннями препозитивних дієприкметникових зворотів, наприклад, українська конструкція … *сформульований^{a} на роздрукованому^{b} для заспокоєних^{c} узгоджений^{d} зі всіма розклад$^{d'}$ викладачів$^{c'}$ папері$^{b'}$ наказ$^{a'}$* …;
2. Слов'янські конструкції пар однорідних іменників, узгоджених за відмінком, наприклад, українська конструкція *моделювання^{a} процесів^{b} опрацювання^{c} інформаційних ресурсів^{d} в системах^{e} електронної контент^{f} – комерції та корпоративних Інтентет$^{f'}$ – порталах, або в бізнес$^{e'}$ – проектах комерційного контенту$^{d'}$, та реалізація$^{c'}$ програмних засобів$^{b'}$, а також їх активне впровадження$^{a'}$* …, або польська конструкція … *metody^{a} realizacji^{b} systemy^{c} elektroniczny content - commerce^{d} i Internet projektów$^{d'}$ lub przetwarzanie$^{c'}$ zasobów$^{b'}$ informacji i narzêdzi programowych$^{a'}$* …[6, 7, 10, 14–16, 19, 20, 32, 33, 36, 45, 48–57];
3. Конструкції романських мов , наприклад, для українського речення *методи реалізації систем електронної контент-комерції та інтернет-проектів, або процесів опрацювання інформаційних ресурсів, а також програмні засоби* є $S'de\,S''de\,S'''...A'''A''A'$ для французької (*…méthodes^{a} de mise^{b} en œuvre de systemes^{c} de commerce^{d} électronique contenu et projets$^{d'}$ Internet, ou du déroulement$^{c'}$ du ressources$^{b'}$ informationnelles et des outils logiciels$^{a'}$*…), $S'of\,S''of\,S'''...A'''A''A'$ для англійської (*… methods^{a} of implementation^{b} of systems^{c} of electronic content commerce^{d} and Internet projects$^{d'}$, or the processing$^{c'}$ of information resources$^{b'}$ and software tools$^{a'}$* …) [1, 34, 58-60, 62-66] та $S'von\,S''von\,S'''...A'''A''A'$ для німецької

(… $Methoden^{a}$ der $Umsetzung^{b}$ von $Systeme^{c}$ von $elektronischen$ $Content$ $Geschäftsverkehr^{d}$ und $Internet$ $Vorhaben^{d'}$ $oder$ die $Verarbeitung^{c'}$ von $Informationsmitteln^{b'}$ und $Softwarewerkzeuge^{a'}$ …) [25, 30-31, 35, 42] мов.

Відомі також ще два приклади непридатності граматики G_3 (у цих випадках непридатні і граматики G_2, які належать до обмежених за поширеністю явищ. Що ж стосується приведених прикладів, то, хоча список мов, в яких припустимі подібні конструкції, ймовірно, обмежений, вони все ж таки є цілком типовими, так що повністю ігнорувати їх не можна. Тому доводиться визнати, що побудувати повний опис природної мови на основі лише граматики G_3 неможливо. Це означає: або побудована граматика G_3 не породжуватиме деяких правильних фраз (зокрема, речень вигляду $abcd\ldots d'c'b'a'$), або, якщо зробимо її здатною породжувати будь-яку правильну фразу, то вона обов'язково почне породжувати і деякі неправильні фрази (наприклад, поряд з реченням $abcd\ldots d'c'b'a'$ породжуватиме речення $abcd\ldots c'a'd'b'$ – з порушеним узгодженням). Надалі, коли йдеться про неможливості описати мову за допомогою тієї або іншої граматики, мають на увазі саме це – або граматика не породжує деяких правильних фраз доволі звичайного і поширеного типу (тобто є неповною), або обов'язково породжує, окрім всіх правильних фраз, і деякі неправильні (тобто є неадекватною).

Оскільки граматики G_3 недостатньо для опису природної мови у всьому об'ємі, ще не можливо описувати за допомогою граматик G_3 ті або інші фрагменти природної мови. При цьому можна, як видається, передбачати, що в природній мові "G_3–фрагмент", як правило, покриває головну частину. Насправді конструкцій з необмеженим вкладенням пар, що погоджуються, небагато більше, а інколи до того ж вони ще і периферійні. Тим самим граматики G_3 в принципі здатні описувати доволі істотну частину множини речень (простих і складних) природної мови. Крім того, граматики G_3 можуть описувати й інші мовні об'єкти: наприклад, словосполучення елементарних іменних груп, словоформи, склади. Зрозуміло, із сказаного зовсім не випливає, що у всіх тих випадках, коли граматики G_3 застосовні, вони описують свій об'єкт природним чином, тобто вони не завжди зручні. Більш того, з попереднього викладу видно, що це не так. Проте будь-яка граматика, спеціально пристосована для природного опису якого-небудь G_3–фрагмента мови (скажімо, для опису простих речень, що не містять конструкцій на зразок вказаних вище), буде еквівалентна деякій граматиці G_3. А оскільки граматики, еквівалентні граматикам G_3, зазвичай в якомусь відношенні – за характером або правил, або виведення – характеризуються приблизно такою ж мірою простоти, що і самі граматики G_3, то тим самим отримуємо ніби еталон простоти. Створюючи граматику, що описує прості речення, повинні прагнути до того, щоби вона хоч би в одному з вказаних відношень не була набагато складнішою за граматику G_3 [14–16].

Що ж до питання про те, де саме граматики G_3 виявляються не лише застосовними, але і природними, то загалом це досі залишається малодослідженим питанням, і з'ясувати його цікаво. Насамперед можна вважати, що граматики G_3 достатньою мірою зручні при описі елементарних іменних груп типу *або не лише зі всіх процесів три перших етапи моєї моделі життєвого циклу комерційного контенту* (тут наведено приклад максимальної *схеми елементарної іменної групи* – *Eng* (elementary nominal group); насправді такі групи зазвичай мають простіший вигляд – ті або інші місця можуть бути не заповнені) [14–16].

Під елементарною іменною групою тут розуміють іменник із всіма його препозитивними узгодженими визначеннями, а також прийменником, (обмежувальними) частками і сурядним сполучником, що вводить усю групу. Наведемо приклад граматики G_3, що породжує будь-які *Eng* вказаного типу (з неживими S).

Схема граматики G_3 [6, 7, 10, 14–16, 19, 20, 32, 33, 36, 45, 48–57].

I. $Eng \to \{або, і, та, чи, ..., \Lambda\} Eng^1$,

де Λ – порожній ланцюжок; змістовна наявність Λ означає, що відповідне місце в Eng може залишитися незаповненим.

II. $Eng^1 \to \{ні, не, лише, тільки, хоча б, ..., \Lambda\} Eng^2$

III. 1. $Eng^2 \to \left\{ \begin{matrix} біля, у, в, від, для, до, за, з, із, без, між, \\ проти, серед, ради, заради, задля, ..., \Lambda \end{matrix} \right\} Eng^3_{x,y,род}$

2. $Eng^2 \to \left\{ \begin{matrix} завдяки, наперекір, назустріч, на перевагу, \\ навперейми, услід, усупереч, ..., \Lambda \end{matrix} \right\} Eng^3_{x,y,дат}$

3. $Eng^2 \to \left\{ \begin{matrix} в, у, з, за, зважаючи на, з огляду на, крізь, \\ над, незважаючи на, о, об, перед, під, по, \\ поміж, понад, попереду, попід, проміж, \\ на, між, поза, попри, через, про, повз, ..., \Lambda \end{matrix} \right\} Eng^3_{x,y,знах}$

4. $Eng^2 \to \left\{ \begin{matrix} услід\ за, з, за, згідно\ з, між, над, перед, \\ у\ зв'язку\ з, під, поза, поміж, понад, \\ попід, порівняно\ з, поруч\ з, поряд\ з, \\ проміж, разом\ з, слідом\ за, ..., \Lambda \end{matrix} \right\} Eng^3_{x,y,оруд}$

5. $Eng^2 \to \{в, у, на, о, об, по, при, ..., \Lambda\} Eng^3_{x,y,місц}$

6. $Eng^2 \to Eng^3_{x,y,z}$

IV. $Eng^3_{x,y,z} \to \{весь_{x,y,z}, кожний_{x,y,z}, який-небудь_{x,y,z}, ..., \Lambda\} Eng^4_{x,y,z}$

V. $Eng^4_{x,y,z} \to \{цей_{x,y,z}, той_{x,y,z}, такий_{x,y,z}, ..., \Lambda\} Eng^5_{x,y,z}$

VI. 1. $Eng^5_{x,y,z} \to \{один_{x,y,z}, \Lambda\} Eng^6_{x,y,z}$

2. $Eng^5_{x,y,z} \to \{два_{x,z}, три_z, чотири_z, \Lambda\} Eng^6_{x,од,род}$

Тут $z = наз, оруд$.

3. Тут $z = наз, оруд$.

$$Eng^5_{x,y,z} \to \left\{ \begin{matrix} п'ять_z, шість_z, сім_z, вісім_z, дев'ять_z, ..., \\ тисяча\ дев'ятсот\ сімдесят\ дев'ять_z, ..., \Lambda \end{matrix} \right\} Eng^6_{x,мн,род}$$

4. $Eng^5_{x,y,z} \to \left\{ \begin{matrix} два_{x,z}, ..., п'ять_z, шість_z, сім_z, вісім_z, ..., \\ тисяча\ дев'ятсот\ сімдесят\ дев'ять_z, ..., \Lambda \end{matrix} \right\} Eng^6_{x,мн,z}$

Тут $z \neq наз, оруд$.

VII. $Eng^6_{x,y,z} \to \{мій_{x,y,z}, твій_{x,y,z}, ..., їх_{x,y,z}, \Lambda\} Eng^7_{x,y,z}$

VIII. $Eng^7_{x,y,z} \to \left\{ \begin{matrix} перший_{x,y,z}, другий_{x,y,z}, ..., \\ тисяча\ дев'ятсот\ сімдесят\ дев'ятий_{x,y,z}, ..., \Lambda \end{matrix} \right\} Eng^{8i}_{(1 \le i \le p)\ x,y,z}$

IX. $Eng^{8i}_{(1 \le i \le p)\ x,y,z} \to a^i_{x,y,z}\ Eng^{8j}_{(i \le j \le p)\ x,y,z} \left(1 \le i \le j \le p\right)$

Позначення $a^i_{x,y,z}$ роз'яснюється в примітці після граматики.

X. $Eng^{8i}_{(1 \le i \le p)\ x,y,z} \to S_{x,y,z}$

XI. 1. $S_{чол,у,z} \to \{тато_{у,z}, дід_{у,z}, стіл_{у,z}, контент_{у,z}, зошит_{у,z}, метод_{у,z}, ...\}$

2. $S_{жін,у,z} \to \{дитина_{у,z}, мама_{у,z}, посмішка_{у,z}, модель_{у,z}, зірка_{у,z}, ...\}$

3. $S_{сер,у,z} \to \{око_{у,z}, щастя_{у,z}, море_{у,z}, серце_{у,z}, сонце_{у,z}, ...\}$

Зауваження до правила IX. Символ a^i використано тут для позначення конкретних прикметників i-го класу, причому до одного класу належать прикметники, що займають одну і ту саму позицію відносно визначуваного іменника, а нерівність $i > j$ означає, що прикметник i-го класу повинен стояти далі від іменника, ніж прикметник j-го класу. Наприклад, прикметник *український* (*англійський*, *німецький* ...) має індекс класу менший, ніж у прикметника *цікавий* (*новий*, *коштовний* ...), оскільки висловлювання *цікаві українські журнали* природніше, ніж *українські цікаві журнали*. Кількість таких класів прикметників позначено через p [14–16].

Отже, до граматики G_3 має бути прикладений список прикметників, забезпечених індексами класу в певному тут значенні. Для прикладу візьмемо невеликий словник, що містить прикметники шести класів (табл. 3) [14–16].

Таблиця 3

Словник класифікації прикметників української мови

№	1-й клас	2-й клас	3-й клас	4-й клас	5-й клас	6-й клас
1	лінгвістичний	інформаційний	англійський	зелений	добрий	загальний
2	математичний	дерев'яний	український	синій	веселий	перший
3	комп'ютерний	металевий	німецький	жовтий	чудовий	уніфікований
4	музичний	кістяний	польський	червоний	цікавий	типовий
5	фізичний	аналітичний	французький	фіолетовий	відмінний	детальний
6	класичний	паперовий	американський	коричневий	спритний	останній
…	.....	.....	.....	.....	.....	.....

Це розбиття виконано винятково в ілюстративних цілях і відображає реальну картину вельми наближено: закони взаємного розміщення прикметників насправді не укладаються в рамки лінійного впорядкування; крім того, порядок прикметників залежить насправді від логічного акценту (актуального розчленовування словосполучення) – так, *сучасний лінгвістичний навчальний посібник* ≈ "*є навчальний посібник з лінгвістики, який недавно виданий*", тоді як *математична цікава книга* ≈ "*є сучасна книга як навчальний посібник, вона з лінгвістики*". Аналогічно, в лінгвістичній роботі, присвяченій іменним конструкціям деякої мови, мова може йти про визначальні іменні конструкції, суб'єктні іменні конструкції тощо, а в роботі про визначальні конструкції зустрінемося швидше з іменними визначальними конструкціями, дієслівними визначальними конструкціями тощо. Описуючи такі випадки, маємо на увазі найбільш нейтральний, звичайний порядок. Наведемо приклад виведення висловлювання …*і тільки зі всіма цими трьома моїми уніфікованими аналітичними лінгвістичними методами* … в граматиці G_3:

Eng

(I) *і* Eng^1

(II) *і тільки* Eng^2

(III.4) *і тільки зі* $Eng^3_{чол,мн,оруд}$

(IV) *і тільки зі всіма* $Eng^4_{чол,мн,оруд}$

(V) *і тільки зі всіма цими* $Eng^5_{чол,мн,оруд}$

(VI.4) *і тільки зі всіма цими трьома* $Eng^6_{чол,мн,оруд}$

(VII) *і тільки зі всіма цими трьома моїми* $Eng^7_{чол,мн,оруд}$

(VIII) *і тільки зі всіма цими трьома моїми* $Eng^{8_5}_{чол,мн,оруд}$

(IX) *і тільки зі всіма цими трьома моїми уніфікованими* $Eng^{8_2}_{чол,мн,оруд}$

(IX) *і тільки зі всіма цими трьома моїми уніфікованими аналітичними* $Eng^{8_1}_{чол,мн,оруд}$

(IX) *і тільки зі всіма цими трьома моїми уніфікованими аналітичними лінгвістичними* $Eng^{8_1}_{чол,мн,оруд}$

(X) *і тільки зі всіма цими трьома моїми уніфікованими аналітичними лінгвістичними* $S_{чол,мн,оруд}$

(XI.1) *і тільки зі всіма цими трьома моїми уніфікованими аналітичними лінгвістичними методами*

Звернемо увагу на те, що складові, які виходять із приведеного виведення способом в табл. 2, виявляються в цьому випадку цілком природними (на відміну від граматики G_2) (див. рис. 2). Це пояснюється особливостями синтаксичної будови елементарних іменних груп в українській мові – *Eng* будуються за схемою

$$...\Big(h\Big(g\Big(f\Big(e\Big(d\Big(c\big(b(a)\big)\Big)\Big)\Big)\Big)\Big)\Big)...,$$

тобто всі елементи, що поширюють головний елемент, знаходяться ліворуч від нього, і кожен з них належить (визначає або підпорядковує) відразу до всього, що стоїть услід за ним; при цьому будь-який елемент, окрім головного (тобто останнього), може бути відсутнім. Допоміжні символи, що зустрічаються у виведенні, – Eng^1, Eng^2, $Eng^3_{чол,мн,оруд}$, ... – природно інтерпретуються як позначення для *неповних* Eng, тобто як типів складових: Eng^1 – іменна група без сполучника, Eng^2 – іменна група без сполучника і без обмежувальної (заперечної) частки, $Eng^3_{чол,мн,оруд}$ – іменна група без сполучника, без обмежувальної (заперечної) частки і без прийменника тощо.

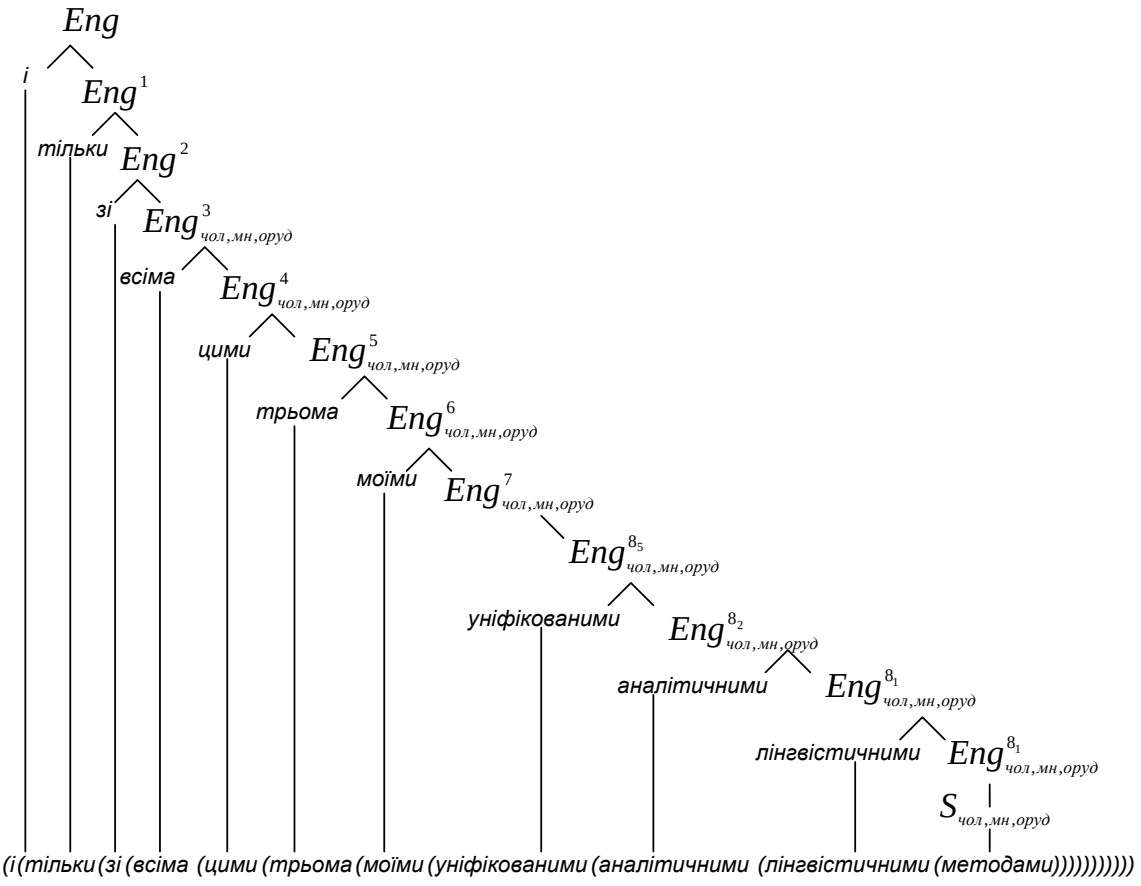


Рис. 2. Приклад 2 регулярної граматики G_3

Взагалі, граматики G_3 в деякому відношенні зручні для опису саме тих мовних об'єктів, які мають вказану схему будови, тобто "нарощуються" лише з одного боку. Такими об'єктами, наприклад, є агглютинативно побудовані словоформи – на зразок таких, як українські $((((контроль)ова)н)ий)$ або $((((контроль)ова)н)ість)$, а також $((((нудьг)у)юч)ий)$ тощо (у мовах типу турецької або угорської подібні форми мають незрівнянно регулярніший характер).

Зрозуміло, для опису об'єктів, зростаючих управо, природніші граматики G_3 не зовсім такі, як визначено в табл. 1, а саме: у цьому визначенні слід замінити правила вигляду $A \to bB$ правилами вигляду $A \to Bb$, тобто перейти від лівобічних граматик G_3 до правобічних.

На закінчення роз'яснимо, що треба розуміти під зручністю застосування граматик G_3 для опису мовних об'єктів вигляду $\Big(\Big(\Big(\Big(\Big(\Big(\Big(\Big((a)b\Big)c\Big)d\Big)e\Big)f\Big)g\Big)h\Big)$ – так би мовити, *шаруватих* об'єктів. Граматики G_3 зручні тут саме тому, що ця шаруватість в явній формі розкривається G_3– виведенням відповідного об'єкта (приклад 2). Проте в інших аспектах граматики G_3 можуть і не бути зручними при описі навіть таких об'єктів, наприклад, для відбиття всіх морфонологічних процесів, супроводжуючих породження словоформ. У таких випадках, мабуть, виявиться доцільним розчленовувати відповідні явища на різні рівні і описувати їх декількома граматиками, одна з яких буде автоматною (наприклад, граматиці G_3 може бути доручене породження словоформ на рівні морфем і, можливо, морф, тоді як подальша реалізація отриманого ланцюжка виконується граматиками інших типів) [14-16].

Відбір контенту з різних джерел інформації. Особливістю сучасності є постійний ріст темпів виробництва контенту. Цей процес є об'єктивним і позитивним, але виникла проблема: прогрес в галузі виробництва контенту призводить до пониження загального рівня інформованості потенційного користувача. Крім збільшення обсягів контенту до масштабів, яке призводить до неможливості його безпосереднього опрацювання, та швидкості його поширення виникає низка специфічних проблем (табл. 4).

Таблиця 3

Основні негативні чинники у формуванні комерційного контенту

Назва	Основна причина	Рішення
Інформаційний шум	Структурованість масивів контенту.	Фільтри, контент-моніторинг, аналіз сайту, контент-аналіз.
Паразитичний контент	Поява як додатків	Фільтри, контент-моніторинг, контент-аналіз.
Нерелевантність контенту	Невідповідність потребам користувачів.	Створення анотованої бази даних, пошукових образів первинного контенту та їх кластеризація, контент-аналіз.
Дублювання контенту	Дублювання в джерелах.	Контент-аналіз, сканери і фільтри на базі статистики та критеріїв.
Навігація в потоці контенту	Швидкий ріст обсягу і поширення контенту.	Аналіз сайту, фільтри, контент-моніторинг, контент-аналіз.
Надмірність пошуку	Дублювання і нерелевантність.	Анотований пошук, контент-аналіз та реферування.

Негативні чинники у формуванні контенту ускладнюють процес пошуку необхідних даних при скануванні різних джерел інформації. Оператор створення комерційного контенту $\alpha_0 : (X, U_C, T) \to C_0$ є відображенням вхідних даних з різних джерел інформації у комерційний контент, який відрізняється від попереднього стану комерційного контенту актуальністю. Збільшення фізичного обсягу та змінна актуальності/динаміки контентних потоків (постійне систематичне та нерегулярне оновлення) призводить до виникнення дублювання, інформаційного шуму та надмірності результатів пошуку контенту. Охоплення та узагальнення великих динамічних потоків контенту, які неперервно генерують в інтернет-джерелах, вимагає якісно нових методів/підходів пошуку як контент-моніторинг (рис. 3).

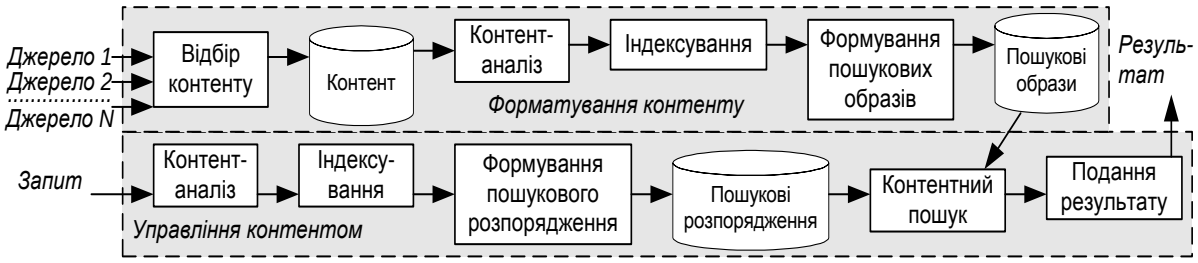


Рис. 3. Структурна схема процесу контент-моніторингу даних $\alpha_1 : (X, U_G, T) \to C_0$

Оператор збирання комерційного контенту $\alpha_1 : (X, U_G, T) \to C_0$ відображенням вхідних даних від авторів або модераторів системи у комерційний контент відрізняється від попереднього стану комерційного контенту достовірністю та актуальністю. Вхідною інформацією для контент-моніторингу є текст природною мовою як послідовність символів, вихідна інформація – це таблиці розділів, речень і лексем аналізованого тексту. Контент-моніторинг є програмним засобом автоматизації знаходження найважливіших складових у потоках контенту. Це змістовний аналіз потоків контенту з метою постійного отримання необхідних якісних/кількісних зрізів протягом наперед не визначеного проміжку часу.

Складовою контент-моніторингу є контентний пошук та контент-аналіз тексту. Контент-аналіз призначений для пошуку контенту в масиві даних за змістовими лінгвістичними одиницями (алг. 1). Одиниця рахунку є кількісною мірою одиниці аналізу, що дає змогу реєструвати частоту (регулярність) появи ознаки категорії аналізу в тексті (кількість певних слів або їх поєднань, рядків, друкованих знаків, сторінок, абзаців, авторських аркушів, площа тексту тощо).

Алгоритм 1. Контент-аналіз текстового комерційного контенту.

Етап 1. Визначення набору критеріїв $< U_C, U_G >$ для текстового контенту X.

Крок 1. Формування набору критеріїв як тип джерела (форум, електронна пошта, інтернет-газета, чат, інтернет-журнал); тип контенту (стаття, е-лист, банер, коментарій); учасники комунікації (відправник, одержувач, реципієнт).

Крок 2. Визначення розміру (мінімальний обсяг або довжина), частоти появи, способу/місця розповсюдження та час появи контенту.

Крок 3. Фільтрування згідно із сформованим набором критеріїв контентного потоку та зберігання ідентифікованого релевантного контенту X.

Етап 2. Контент-аналітичний відбір. Формування вибіркової сукупності контенту X' за критеріями обмеженої вибірки $< U_C, U_G >$ з більшого масиву $X = \{Sentence_1, Sentence_2, ...Sentence_n\}$, де $Sentence_i \to \tilde{S}_{x,y,наз,w} \tilde{V}_{y,тепер,w}$.

Етап 3. Виявлення змістовних одиниць аналізу $< U'_C, U'_G >$ текстового комерційного контенту X' (словосполучення, речення, тема, ідея, автор, персонаж, соціальна ситуація, частина тексту, кластеризована за змістом категорії аналізу). Вимоги до вибору лінгвістичної одиниці аналізу: велика для інтерпретації значення; мала, щоб не інтерпретувати багато значень; легко ідентифікується; кількість одиниць велика для проведення вибірки.

Етап 4. Виділення одиниць рахунку аналізу текстового контенту X'.

Крок 1. Якщо одиниці рахунку $< U_C, U_G >$ збігаються з одиницями аналізу $< U'_C, U'_G >$, то знаходять частоти появи виділеної змістовної одиниці, інакше перейти до кроку 2.

Крок 2. Модератор на основі аналізованого контенту висуває та доповнює одиниці рахунку $< U_C, U_G >$, наприклад, протяжність текстів; площа тексту, заповнена змістовними одиницями; кількість рядків (абзаців, знаків, колонок тексту); розмір/вид файла; кількість рисунків з певним змістом/сюжетом тощо.

Етап 5. Порівняння змістовних одиниць аналізу $< U'_C, U'_G >$ з одиницями $< U_C, U_G >$.

Крок 1. Класифікація за угрупованнями із оціненням ваги змістовних категорій у загальному обсязі тексту. Класифікатором є загальна таблиця, до якої зведено всі категорії аналізу і одиниці аналізу. Фіксують одиниці виразу категорій.

Крок 2. Статистичні розрахунки зрозумілості та атрактивності контенту.

Етап 6. Розроблення інструменту контент-аналізу.

Крок 1. Створення закодованого протоколу контенту X' для компактності подання даних та швидкого порівняння результатів аналізу різного контенту.

Крок 2. Заповнення протоколу контенту X' властивостями (автор, час видання, обсяг тощо).

Крок 3. Заповнення протоколу контенту X' підсумками його аналізу (кількість вживання в ньому певних одиниць аналізу і висновки щодо категорій аналізу). Протокол кожного контенту X' заповнюється на основі підрахунку даних всіх його реєстраційних карток.

Етап 7. Розроблення таблиці контент-аналізу. Тип таблиці визначають у вигляді системи скоординованих і субординованих категорій аналізу: кожна категорія (питання) передбачає ряд ознак (відповідей), за якими квантифікується зміст тексту X'.

Етап 8. Розроблення кодувальної матриці контент-аналізу.

Крок 1. Якщо обсяг вибірки ≥ 100 одиниць, то аналізується набір матричних листів, інакше виконати крок 2.

Крок 2. Якщо вибірка < 100 одиниць, то проводиться двовимірний аналіз. У цьому випадку для кожного контенту X' формується кодувальна матриця.

Етап 9. Проведення аналізу тексту X' згідно із створеними кодувальними матрицями.

Етап 10. Інтерпретація результатів $\alpha_0 : (X, U_C, T) \to C_0$ та $\alpha_1 : (X, U_G, T) \to C_0$. Виявляють і оцінюють характеристики контенту X' на основі статистичного набору підрахованих коефіцієнтів за певний період часу на визначену категорію. Охоплює всі здобуті фрагменти тексту C_0, висновки ґрунтуються не на частині результатів, а враховуються всі без винятку.

Застосування контент-аналізу при моніторингу інтернет-джерел даних дає змогу автоматизувати процес знаходження найважливіших складових у потоках контенту X під час відбору даних з цих джерел. Це усуває дублювання контенту C_0, інформаційний шум, паразитичний контент, надмірність результатів пошуку тощо. Цей метод застосовують на подальших етапах формування контенту C_0 для отримання точнішого релевантного результату – створення унікального контенту C_0, який користується попитом серед користувачів СЕКК.

Процес виявлення дублювання змісту комерційного контенту. Оператор виявлення дублювання комерційного контенту $\alpha_2 : (C_0, T, U_B) \to C_1$ є відображенням комерційного контенту C_0 в новий стан, який відмінний від попереднього стану унікальністю (рис. 4).

Рис. 4. Діаграма станів для процесу виявлення дублювання контенту C_0 в СЕКК

Дубльований текст краще визначати, враховуючи відстань Левенштейна (також функція Левенштейна, алгоритм Левенштейна або відстань редагування) – у теорії інформації і комп'ютерній лінгвістиці міра відмінності двох послідовностей символів (рядків). Обчислюється як мінімальна кількість операцій вставки, видалення і заміни, необхідних для перетворення однієї послідовності на іншу. Щоб перетворити слово “небо” на слово “треба”, необхідно зробити дві заміни та одну вставку, відповідно дистанція Левенштейна становить 3 {*небо* → *неба*, *неба* → *реба*, *реба* → ***т****реба*}. На практиці дистанція Левенштейна використовується для визначення подібності послідовностей символів, наприклад, для корегування орфографії або для пошуку дублікатів.

Для розрахунку відстані Левенштейна найчастіше використовують простий алгоритм, в якому використовується матриця розміром $(n+1)(m+1)$, де n і m – довжини порівнюваних рядків. Окрім цього, вартість операцій вилучення, заміни та вставки вважається однаковою. Для конструювання матриці використовують таке рекурсивне рівняння:

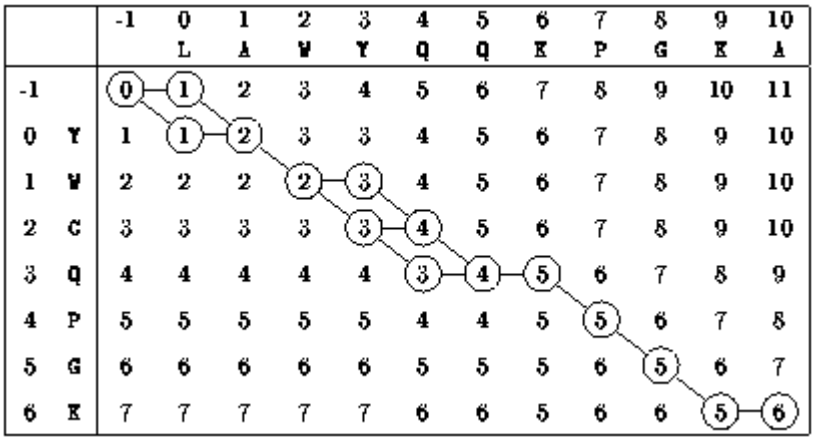
$$D_{0,0} = 0 \text{ та } D_{i,j} = \min\begin{cases} D_{i-1,j-1} + 0 \ (equal, nochange) \\ D_{l-1,j-1} + 1 \ (replace) \\ D_{i-1,j} + 1 \quad (insert) \\ D_{i,j-1} + 1 \quad (delete) \end{cases}$$

Наприклад : x = YHCQPGK і y = LAHYQQKPGKA

		-1	0	1	2	3	4	5	6	7	8	9	10
			L	A	H	Y	Q	Q	K	P	G	K	A
-1		0	1	2	3	4	5	6	7	8	9	10	11
0	Y	1	1	2	3	3	4	5	6	7	8	9	10
1	H	2	2	2	2	3	4	5	6	7	8	9	10
2	C	3	3	3	3	3	4	5	6	7	8	9	10
3	Q	4	4	4	4	4	3	4	5	6	7	8	9
4	P	5	5	5	5	5	4	4	5	5	6	7	8
5	G	6	6	6	6	6	5	5	5	6	5	6	7
6	K	7	7	7	7	7	6	6	5	6	6	5	6

```
( Y — H C Q — — P G K — )
( L A H Y Q Q K P G K A )

( Y — H C — Q — P G K — )
( L A H Y Q Q K P G K A )

( Y — H — C Q — P G K — )
( L A H Y Q Q K P G K A )

( — Y H C Q — — P G K — )
( L A H Y Q Q K P G K A )

( — Y H C — Q — P G K — )
( L A H Y Q Q K P G K A )

( — Y H — C Q — P G K — )
( L A H Y Q Q K P G K A )
```

Це дає шість наступних вирівнювань, які відповідають шести мінімальним витратам шляхів між (-1, -1) і (6,10):



Для визначення послідовності операцій, необхідних для переходу від одного слова до іншого, потрібно знайти найкоротший шлях від першої [0,0] клітинки матриці до останньої [i,j]. Як видно з прикладу, існує декілька еквівалентних шляхів, і алгоритм не тільки вираховує мінімальну відстань,

але й усі шляхи, використовуючи у кожному наступному кроці інформацію, здобуту за попередні кроки (принцип динамічного програмування).

Для відстані Левенштейна існують такі верхня і нижня межі:

1. Дистанція Левенштейна не є меншою за різницю довжини рядків, що порівнюються.
2. Вона не є більшою за довжину найдовшого рядка.
3. Вона дорівнює 0 тоді і тільки тоді, коли рядки однакові (однакові символи на однакових позиціях).

Між відстанню Левенштайна та відстанню Гемінга існують такі взаємозв'язки:

- Для рядків однакової довжини відстань Левенштайна дорівнює відстані Гемінга, оскільки відстань Гемінга користується лише операцією заміни одного символу на інший і не дозволяє вставки та вилучення символів.
- Якщо рядки різної довжини, то верхньою межею є відстань Гемінга плюс різниця довжини рядків.

Усунення дубльованого контенту $C_1 = \alpha_2(\alpha_1(X, U_G, T), U_B)$ в контентних потоках необхідне не завжди. Існують задачі, у яких використовують факт дублювання комерційного контенту C_0 із різних джерел, наприклад, при визначенні його важливості (якщо є дублювання X в різних джерелах) або ефективності PR-кампаній (підрахунок републікацій прес-релізів тощо). При визначенні дублюванні контенту C_0 (алг. 2) досліджують рівень його появи на інформаційних ресурсах, які мають посилання на інші ресурси.

Алгоритм 2. Визначення дублювання контенту.

Етап 1. Завдання початкових умов U_B для визначення дублів c_{0j} з множини C_0.

Крок 1. Завдання модератором кількості слів у ланцюжку $m = const$.

Крок 2. Завдання коефіцієнта унікальності ланцюжків $U = const$.

Крок 3. Завдання меж коефіцієнта появи слів $K = [a_1, a_2]$, де $a_1 = const$ та $a_2 = const$.

Крок 4. Розбиття контенту на n ланцюжків по m слів.

Крок 5. Розрахунок частот появи ключових слів k_i.

Етап 2. Визначення дублів контенту $\alpha_2 : (C_0, T, U_B) \to C_1$.

Крок 1. Якщо множина C_0 порожня, зупинити алгоритм.

Крок 1. Порівняння між собою ланцюжків слів для всього контенту $c_{0j} \in C_0$.

Крок 2. Розрахунок коефіцієнтів унікальності ланцюжків u_i для $c_{0j} \in C_0$.

Крок 3. Порівняння коефіцієнтів унікальності ланцюжків u_i із коефіцієнтом U. При $\frac{1}{n}\sum_{i}^{n} u_i < U$ контент $c_{0j} \in C_0$ маркувати непридатним. Перейти до кроку 1.

Крок 4. Порівняння частоти k_i із коефіцієнтом K. Якщо $k_i < a_1$ або $k_i > a_2$, то контент c_{0j} маркувати непридатним. Перейти до кроку 1.

Етап 3. Записати c_{0j} як $c_{1l} \in C_1$ та вилучити його з C_0. Перейти до етапу 1.

Дубльований за змістом текстовий комерційний контент в СЕКК являють на основі лінгвостатистичних методів, що полягають у виявленні в різному контенті загальних термів, ланцюжки яких творять словесні сигнатури текстового комерційного контенту.

Форматування комерційного контенту. Процес форматування контенту C_1 реалізує вручну модератор або автоматично оператор $\alpha_3 : (C_1, U_{FR}, T) \to C_2$ з використанням семантичних засобів: інформаційно-пошукової мови, методів пошуку та індексування контенту/запитів (алг. 3). Назви текстового комерційного контенту C_1 розкривають їх тему і предмет, але за назвою не ідентифікують контент. Модератор суб'єктивно індексує контент, тому автоматизація

форматування $C_2 = \alpha_3(\alpha_2(C_0, U_B), U_{FR})$ забезпечує уніфікацію контенту C_1 та зменшення витрат на утримання додаткового персоналу.

Алгоритм 3. Форматування текстового комерційного контенту.

Етап 1. Індексування текстового комерційного контенту C_1.

Етап 2. Лінгвістичний аналіз текстового комерційного контенту C_1 (алг. 1).

Етап 3. Визначення основного змісту текстового комерційного контенту C_1.

Етап 4. Відокремлення центральної теми текстового комерційного контенту C_1.

Етап 5. Опис текстового контенту C_1 в термінах інформаційно-пошукової мови.

Етап 6. Перетворення текстового контенту на XML-формат $\alpha_3 : (C_1, U_{FR}, T) \to C_2$.

Крок 1. Визначення заголовку комерційного контенту C_1 в тегах head та title.

Крок 2. Визначення тіла комерційного контенту C_1 в тегах body та html.

Крок 3. Визначення автора комерційного контенту C_1 у відповідних тегах.

Крок 4. Визначення дати створення комерційного контенту C_1 у відповідних тегах.

Крок 5. Визначення додаткових властивостей контенту C_1 у відповідних тегах.

Етап 7. Завантаження шаблону комерційного контенту C_2 (рис. 5).

Етап 8. Запис в шаблон C_2 даних про комерційний контент C_1 в форматі XML.

Етап 9. Збереження відформатованого комерційного контенту C_2 в БД СЕКК.

На рис. 5 подано шаблон комерційного контенту C_2 у форматі XML.

Рис. 5. Структурна схема шаблону відформатованого комерційного контенту C_2

Висновки та перспективи подальших наукових розвідок

В інформаційно-пошуковій мові серед основних елементів не використовують характерні для природної синоніми/омоніми через семантичну неоднозначність. Переваги використання інформаційно-пошукових мов залежать від призначення підсистеми форматування, рівня її оснащення технічними засобами та автоматизації інформаційних процедур і ланки управління. Із вирішенням завдання морфологічного аналізу роботу з текстом реалізують у межах речень природною мовою або з усім текстом як набором лінійно впорядкованих слів, словосполучень і речень. При використанні варіації статистичних методів аналізу ігнорують лінгвістичну

взаємопов'язаність і нелінійність природної мови. Незадіяність проміжних рівнів подання тексту, особливо у вигляді семантичних структур, пояснюється відсутністю ефективних формалізмів опису структури тексту комерційного контенту. При форматуванні контенту C_2 у форматі XML більшість тегів є незаповненими – вони заповнюються на наступних етапах формування комерційного контенту C. Цей етап лише дозволяє привести до єдиного вигляду досліджуваний контент при заповнення шаблону для полегшення роботи з ним надалі.

1. Английская грамматика в доступном изложении // Режим доступу: http://realenglish.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://gasteach.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 // Пер. с фр. И. А. Мельчука под ред. 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.corneillemoliere.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". – 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" – 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. 67. Varga D. Yngve's hypothesis and some problems of the mechanical analysis / D. Varga // Computational Linguistics. – III. – 1964. – P. 47–74.