Ганна Ситар

УДК 81'367:81'373.7: 81'32

АЛГОРИТМ РОЗПІЗНАВАННЯ СИНТАКСИЧНИХ ФРАЗЕОЛОГІЗМІВ У КОРПУСІ ТЕКСТІВ: СПРОБА СТВОРЕННЯ

Статтю присвячено спробі створення алгоритму розпізнавання синтаксичних фразеологізмів української мови. Спираючись на результати попередніх теоретичних досліджень, виділено низку формальних і статистичних критеріїв, які дають змогу відмежувати синтаксичні фразеологізми від інших мовних одиниць у корпусі українськомовних текстів.

Запропоновано блок-схему розпізнавання синтаксичних фразеологізмів, яка об'єднує в собі дві гілки, побудовані відповідно для речень із одночленним та речень із двочленним стрижневим компонентом.

Ключові слова: алгоритм, корпус текстів, модель речення, синтаксичний фразеологізм, українська мова, фразеологізоване речення.

Синтаксичні фразеологізми – особливий тип речення, у якому постійний і змінний компоненти пов'язані ідіоматично і розташовані переважно фіксовано, граматичні зв'язки і прямі лексичні значення слів послаблені (Velichko; Vsevolodova, and Yon Lim Su; "Russkaya Grammatika"; Shmelev; Shvedova; Sytar "Status Syntaksychnykh" та ін.), наприклад: *Що за дитина! Яка дівчина! Чим не привід для хвилювання?! День як день. Кохання є кохання. Разом так разом. Теж мені зробили! Яке там домовились! Зрада і в Африці зрада* та ін. Такі речення належать до дієвих засобів експресивного синтаксису, за допомогою яких мовець виражає ставлення до певного предмета або явища дійсності, висловлює оцінку певної реалії або ситуації, прагне досягнути впливу на почуття адресата мовлення (слухача або читача), сформувати потрібну думку або викликати відповідну дію та ін.[1]

В українському мовознавстві проаналізовано структурні й семантичні ознаки фразеологізованих речень, виділено ступені фразеологізації речень (Lychuk; Lychuk, and Shynkaruk та ін.). Автоматичний аналіз синтаксичних фразеологізмів досі не здійснювався, очевидно, через низку причин. По-перше, типово фразеологічні одиниці (маємо на увазі традиційні, або «лексичні» фразеологізми типу *бити байдики, пекти раків*) вносяться до електронних словників програм автоматичного аналізу тексту або автоматизованого перекладу в готовому вигляді (при цьому можна зафіксувати варіанти фразеологізму, проте передбачити всі можливості їх модифікацій і трансформацій досить складно). У випадку із синтаксичними фразеологізмами такий підхід не є ефективним, оскільки будь-який синтаксичний фразеологізм складається з двох компонентів – стрижневого (незмінного) та змінного. Розглянемо такі речення:

– Лікарю, мені можна буде їсти сало?

– Яке там сало!

– Ну, не зараз, а в майбутньому?

– Яке там майбутнє! (Найкращі анекдоти. Сміємося по-українськи: http://www.cmix.in.ua/category/korotki-i-dotepni-anekdoti/page/4);

Мій батько – хай над ним земля пером! – не брехав і синові не звелів. Та й скажіть мені, будь ласка, яке там добро з тієї брехні? Чували, може, й ви, що брехнею світ перейдеш, та назад не вернешся (Марко Вовчок); *Пророкування, визначення долі як окремих осіб, так і цілих держав, астрологічні прогнози стали улюбленою темою газет і телебачення. Яке там вивчення механізмів, ефективності – настав солодкий час «чосу» – обдирання обивателів, котрі на тлі загальної зневіри й розчарування вже остаточно втратили віру в офіційну медицину…* (Дзеркало тижня. – 7.04.2001 (№ 14 (338))); *А на ранок ви прокинулися зі страшним головним болем, глянули на не нафарбовану, не зачесану сплячу дівчину із животиком і... взагалі вона похропує уві сні! Яке там люблю, вам би тепер втекти чимпошвидше* (Час і Події. – 10.08.2009); *…на факультеті належиться всміхатись і на всі "How are you doing?" відповідати "Fine", – це ще одне з арифметичних правил, хоч яке там у лиха "файн", де воно є, те "файн", і хто його бачив* (Оксана Забужко).

У наведених реченнях стрижневим є компонент *яке там*, а змінним – *сало, майбутнє, добро* (із залежними компонентами *з тієї брехні*), *вивчення* (із залежними компонентами *механізмів, ефективності*); *люблю*, *файн.*

Змінний компонент синтаксичних фразеологізмів часто практично не має обмежень щодо лексичного наповнення, характеризується наявністю морфологічних варіантів реалізації (див. попередні приклади) та значними можливостями граматичного розширення за рахунок реалізації його синтаксичних зв'язків. Так,

[1] Цей своєрідний тип речень на матеріалі різних мов привертав увагу представників різних лінгвістичних шкіл та напрямів: передусім, широкого підходу до фразеології (В. Л. Архангельський, А. М. Баранов, Д. О. Добровольський та ін.), розмовного синтаксису (Н. Ю. Шведова, Д. М. Шмєльов та ін.), функційно-комунікативного синтаксису (М. В. Всеволодова, А. В. Величко, Л. О. Балобанова та ін.), Московської семантичної школи (І. О. Мельчук, Ю. Д. Апресян, Л. Л. Іомдін, Л. М. Іорданська, В. Ю. Апресян та ін.), конструкційної граматики (Ч. Філлмор, П. Кей, А. Голдберг, М. Фрайд та ін.), у межах яких дослідники оперують термінами, усталеними в межах відповідної концепції – «стійка фраза», «фразеологізоване речення», «синтаксичний фразеологізм», «комунікативний фразеологізм», «синтаксична ідіома», «граматична ідіома», «фразеосхема», «фразема», «конструкція малого синтаксису», «конструкція мікросинтаксису», «нестандартна конструкція», «формальна ідіома» («лексично вільна ідіома») і под.

фразеологізована модель речення *Що за* N_1 Cop_f може мати реалізації з атрибутивними, об'єктними та суб'єктними поширювачами: ... *Що за місце якесь недобре! Я то казав колись Михайлови – не доточуй стріху ззаду, поклади веранду, як хочеш розширити хату, або комору вибудуй збоку, але стріху ззаду не доточуй…*(Марія Матіос. Солодка Даруся); *Але що за свято без зірок сучасної рок-музики. Цього року хед–лайнером фестивалю став легендарний «Вій»* (Україна молода. – 25.07.2009 (№ 134)); *Але що за смішний був у нас Трансвааль, то я й досі не можу забути. Ото наробили шелесту!* (Іван Нечуй-Левицький); – *І що за диво, що за людина пан Серединський! Ох, мати божа! – аж вищала Теодозя, склавши долоні і дивлячись на стелю* (Іван Нечуй-Левицький) і под.

По-друге, суто формальних ознак синтаксичних фразеологізмів замало, щоб ідентифікувати їх у тексті, водночас змістовий аспект і прагматичне навантаження цих специфічних одиниць майже не піддаються формалізації. Зазначимо, що поодинокі спроби створення алгоритмів виявлення стійких сполук (не синтаксичних фразеологізмів, а стійких сполук у широкому розумінні цього терміна) спираються передусім на кількісні критерії, а саме відношення частоти біграми – аналізованого сполучення двох слів – до частоти триграми – поєднання біграми і слова зліва та поєднання біграми і слова справа (Gusev). Зокрема, на матеріалі російської мови запропоновано кваліфікувати ланцюжок слів стійким, «якщо кількість усіх можливих різних його контекстів (як ліво-, так і правобічних) є достатньо великою, наприклад, зіставлюваною з частотою вживання цього ланцюжка», а зазначені відношення частот ≤ 0,5 (Gusev). При цьому під ланцюжком слів мають на увазі два і більше слів, розташованих підряд у тексті та не перерваних знаками пунктуації, без урахування наявності синтаксичного зв'язку між ними.

По-третє, коректний автоматичний аналіз речень може бути здійснений тільки в межах репрезентативного корпусу текстів, проіндексованого за низкою морфологічних та інших параметрів. Відповідно створити лінгвістичний алгоритм, програмно реалізувати його та здійснити подальшу корекцію програми можуть тільки дослідники, які працюють у межах проектів із створення корпусу текстів. Таких проектів в Україні відомо два – Корпус українських текстів, створений співробітниками лабораторії комп'ютерної лінгвістики Інституту філології Київського національного університету імені Тараса Шевченка під керівництвом Н. П. Дарчук (http://www.mova.info/corpus.aspx?l1=209) та Український національний лінгвістичний корпус (далі УНЛК), розроблений колективом Українського мовно-інформаційного фонду НАН України на чолі з В. А. Широковим (http://unlc.icybcluster.org.ua/virt_unlc/). Описаний нижче алгоритм аналізу синтаксичних фразеологізмів плануємо на подальшому етапі дослідження програмно реалізувати на основі УНЛК.

Вважаємо, що сама постановка завдання автоматичного аналізу синтаксичних фразеологізмів має теоретичне значення, оскільки дає змогу з'ясувати, чи можуть опрацьовані в науковій літературі властивості цих специфічних одиниць, що перебувають на межі синтаксису і фразеології, бути достатніми для їхньої ідентифікації в тексті. Спроба побудови відповідного алгоритму може бути корисною для вдосконалення процедур автоматичного опрацювання тексту та створення систем автоматизованого перекладу, що й зумовлює актуальність пропонованого дослідження.

Мета цього дослідження – побудувати алгоритм розпізнавання в корпусі текстів синтаксичних фразеологізмів із одно- або двочленним стрижневим компонентом. Поставлена мета передбачає розв'язання таких завдань: 1) виділити властивості синтаксичних фразеологізмів, не тільки релевантні для відмежування їх від інших мовних одиниць, але й придатні для формалізації; 2) створити алгоритм розпізнавання в тексті синтаксичних фразеологізмів на матеріалі української мови з урахуванням виділених ознак та різних типів цих одиниць; 3) подати пояснення кроків запропонованого алгоритму.

Аналіз фактичного матеріалу, дібраного з текстів української художньої літератури кінця XIX – початку XXI ст., періодичних видань і українськомовних інтернет-ресурсів, обсягом більше 5000 контекстів уживання синтаксичних фразеологізмів дає змогу виділити дві основні групи цих одиниць[2].

1. Першу групу становлять фразеологізовані речення, що базуються на повторі компонентів, який «цементує» речення. При цьому повтор може реалізуватися у двох підтипах.

1.1. Речення із суміжним повтором, за якого повторюється та сама форма слова, тобто наявний повний збіг лексем, між якими, за нашими спостереженнями, вживаються стрижневі компоненти *є, це, то, не, так, як, чи не*: *Завтра є завтра; Америка – це Америка; Москва – то Москва; Діти як діти. Працювати так працювати*. У цій групі кількісно переважають іменникові речення, можливі дієслівні, прислівникові, займенникові, прикметникові, рідко числівникові речення (Sitar).

1.2. Речення із контактним повтором, у яких вжито різні словоформи одного слова. До цієї групи належать речення з повторюваними іменниками у різних відмінкових формах. Якщо перше і третє слово є словоформами одного слова, то між ними в нашому матеріалі зафіксовано прийменники *над, з/із/зі, між, серед*: *Пан над панами. Красуня з красунь* і под. Якщо різні форми слова виступають першим та другим компонентом, то після них стоїть кома та вживається один із таких сполучників: *а, але, зате, проте, однак*: *Робота роботою, а про здоров'я треба піклуватись; Канікули канікулами, але читати треба.*

[2] Основні структурні й семантичні типи синтаксичних фразеологізмів диференційовано у праці (Sytar "Strukturni y Semantychni Tury"), проте для створення алгоритму розмежування цих конструкцій у тексті виявилось важливим протиставлення речень із повторами та без них.

2. Другу групу утворюють синтаксичні фразеологізми, стрижневий компонент яких складається з поєднання кількох лексем – службових і повнозначних слів, останнім (типово займенникам та прислівникам) властиве семантичне спустошення або семантичний зсув: *не до, от тобі/вам і/й, теж мені, чим не, що за, яке там* і под.: *Не до свята зараз нам. От вам і пісня! Теж мені відмінник! Чим не подарунок?! Що за відповідь! Яке там відпочили!*

На підставі аналізу речень цієї групи укладено реєстр моделей синтаксичних фразеологізмів із двочленними стрижневими компонентами, який на сьогодні охоплює 39 моделей. Із урахуванням можливих родових та числових варіантів відповідних лексем (наприклад, *який там / яка там / яке там / які там*), чергування голосних (наприклад, *ну й / ну і*), наявності загальнолітературного інваріанта і розмовного варіанта (зокрема, *чому не / чом не*) реєстр стрижневих двочленних компонентів включає 92 складені одиниці.

Таблиця 1

Реєстр двочленних стрижневих компонентів синтаксичних фразеологізмів в українській мові

№ з/п	Стрижневий компонент синтаксичного фразеологізму
1.	*ати-бати, йшли*
2.	*ач яка*
3.	*ач яке*
4.	*ач який*
5.	*ач які*
6.	*буду я*
7.	*де вже*
8.	*де вам*
9.	*де йому*
10.	*де їй*
11.	*де їм*
12.	*де мені*
13.	*де тобі*
14.	*де нам*
15.	*де там*
16.	*до чого*
17.	*куди вам*
18.	*куди вже*
19.	*куди йому*
20.	*куди їй*
21.	*куди їм*
22.	*куди мені*
23.	*куди нам*
24.	*куди тобі*
25.	*куди там*
26.	*не до*
27.	*ну і*
28.	*ну й*
29.	*ось так*

30.	*ось … так*
31.	*ось і*
32.	*ось вам*
33.	*ось тобі*
34.	*ось яка*
35.	*ось яке*
36.	*ось який*
37.	*ось які*
38.	*от і*
39.	*от так*
40.	*от … так*
41.	*от вам*
42.	*от тобі*
43.	*ото вам*
44.	*ото тобі*
45.	*от яка*
46.	*от яке*
47.	*от який*
48.	*от які*
49.	*оце так*
50.	*оце … так*
51.	*оце вже*
52.	*оце ж*
53.	*оце і*
54.	*оце й*
55.	*оце вам*
56.	*оце тобі*
57.	*така вже*
58.	*така ж*
59.	*таке вже*
60.	*таке ж*
61.	*такий вже*
62.	*такий уже*
63.	*такий же*
64.	*такі вже*
65.	*такі ж*
66.	*теж мені*
67.	*чи до*
68.	*чи не*
69.	*чим не*

70.	*чом не*
71.	*чому не*
72.	*чому… не*
73.	*що за*
74.	*як не*
75.	*яка вже*
76.	*яка ж*
77.	*яка там*
78.	*яке вже*
79.	*яке ж*
80.	*яке там*
81.	*який вже*
82.	*який же*
83.	*який там*
84.	*який уже*
85.	*які вже*
86.	*які ж*
87.	*які уже*
88.	*які там*
89.	*не винен*
90.	*не винні*
91.	*не пахне*
92.	*не пахнуть*

Зазначимо, що всі стрижневі компоненти, до складу яких входять дієслова або прикметники, функціонують у реченнях, що містять прецедентний вислів, який зазнає трансформації, своєрідного «розхитування» внаслідок різного лексичного наповнення моделі, пор.: *Коні не винні. Бджоли не винні. Яйця не винні. Мораторій не винен* і под. (докладніше див. (Sytar "Syntaksychni Frazeolohizmy")). Належність таких одиниць до синтаксичних фразеологізмів не піддається формалізації і встановлюється на підставі окремого аналізу вилучених контекстів.

Крім перерахованих вище параметрів, до автоматичного аналізу синтаксичних фразеологізмів залучено статистичний критерій – показник асоціації MI (від англ. mutual information – взаємна, спільна, повна інформація) – коефіцієнт, який відбиває невипадковість (залежність) певної послідовності слів у тексті (Church). Для двокомпонентних конструкцій (біграм) обчислення здійснювали за формулою:

(1)

$$\mathrm{MI}\,(\mathrm{x},\mathrm{y}) = log_2 \frac{f(\mathrm{x},\mathrm{y})\times N}{f(\mathrm{x})\times f(\mathrm{y})},$$

де *MI* – коефіцієнт mutual information;

x – перша лексична одиниця;

y – друга лексична одиниця;

f(x,y) – абсолютна частота вживання біграми *xy* в корпусі (з урахуванням порядку одиниць усередині біграми);

f(x) – абсолютна частота *x* в корпусі;

f(y) – абсолютна частота *y* в корпусі;

N – загальна кількість словоформ у корпусі;

$\log_2$ – логарифм числа за основою 2.

Запропоновану процедуру статистичного аналізу синтаксичних фразеологізмів української мови та переваги коефіцієнта MI порівняно з іншими показниками асоціації показано у статті (Sytar "Statystychni

Kryteriyi Analizu"). Здійснені обчислення демонструють, що для синтаксичних фразеологізмів із двочленним незмінним компонентом коефіцієнт MI перебуває в межах від 9,41 (*от яке*) до 15,17 (*ось тобі*). Відповідно одним із кроків алгоритму є встановлення MI для конструкцій, що претендують на статус синтаксичних фразеологізмів. До останніх зараховуємо ті конструкції, для яких обчислене MI ≥ 9.

Зазначимо, що для групи речень із повторами також було б цікаво і перспективно обчислити показники асоціації, проте на матеріалі УНЛК зробити це не видається за можливе, оскільки форма запитів передбачає пошук конкретних слів чи їх поєднань, а не морфологічних класів слів чи певних форм без лексичного наповнення (N_1, N_5, Inf і под.), тобто у корпусі поки що не реалізовано можливість здійснення запитів типу Inf *так* Inf або N_1, N_5, *а* та отримання відповідних частот.

Графічно створений алгоритм втілено у вигляді блок-схеми (див. схему 1). Відповідно до Державного стандарту 19.701-90, який регламентує умовні позначки і правила виконання алгоритмів, програм, даних і систем (GOST 19.701-90), термінатор позначає початок та кінець роботи алгоритму; прямокутник – процес, тобто виконання однієї або кількох операцій; ромб – рішення або функцію перемикального типу з одним входом і двома альтернативними виходами, з яких тільки один може бути активований після виконання вказаних умов.

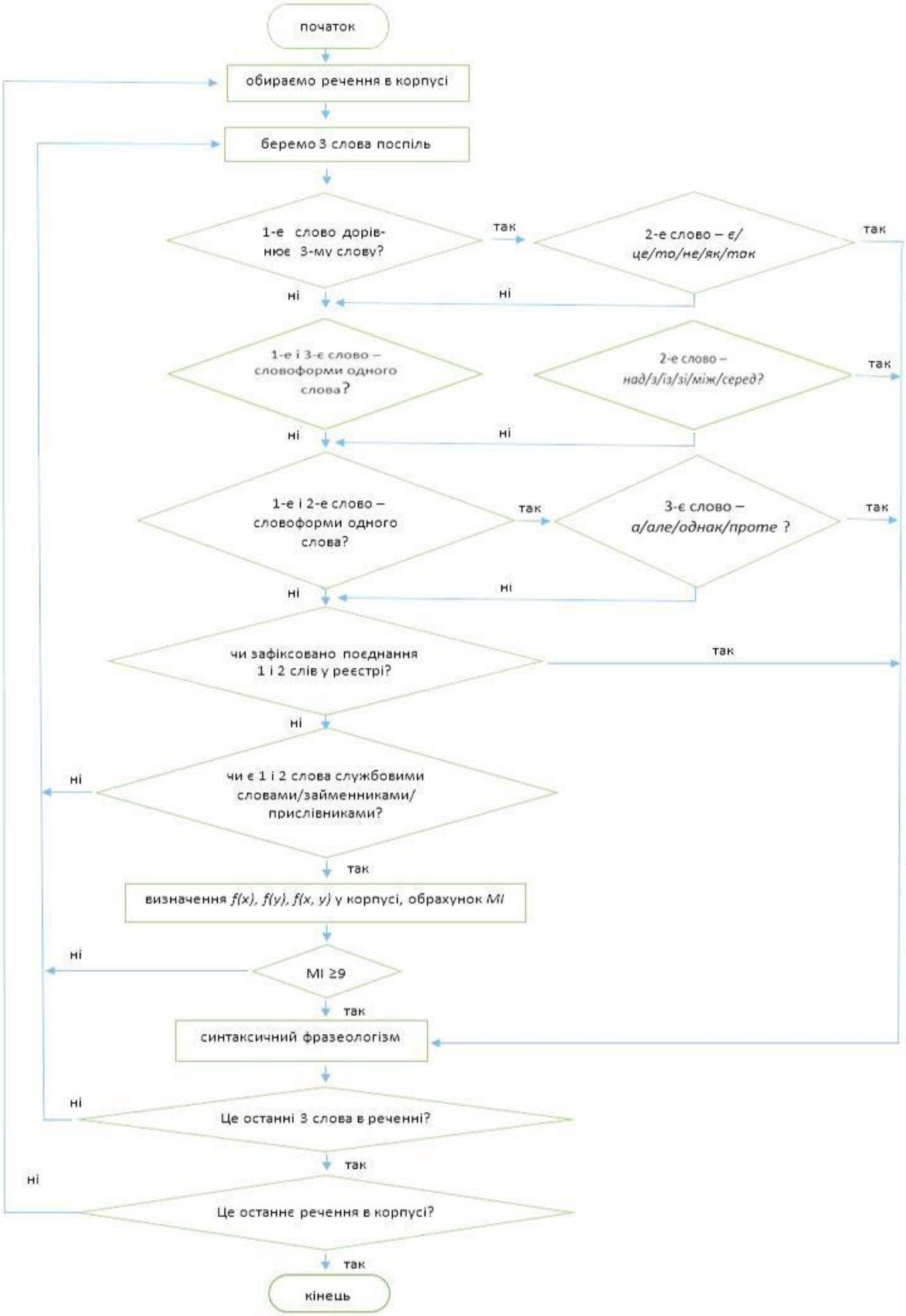
Через складність розв'язання поставлених завдань та потребу врахування комплексу ознак фразеологізованих речень та наявність різних типів цих одиниць запропонований алгоритм має досить складну структуру. Він містить усі три типи базових структур алгоритмів – слідування, розгалуження та повторення. Відповідно цей алгоритм поєднує в собі ознаки лінійного, розгалуженого та циклічного алгоритмів. Розглянемо його докладніше.

1. На вході алгоритму маємо корпус текстів, з якого вичленовуємо речення. Під реченнями в суто формальному плані розуміємо послідовності кількох слів, відокремлені одна від одної такими знаками пунктуації, як крапка, знак оклику, знак питання, знак питання і знак оклику, три крапки або крапка з комою.
2. Оскільки синтаксичні фразеологізми є багатокомпонентними утвореннями, що містять три і більше компоненти, у межах кожного речення аналізуємо по три слова поспіль (усі можливі лінійно розташовані поєднання трьох слів).
3. Далі перевіряємо вичленовані три слова на наявність повторів. Спочатку з'ясовуємо, чи реалізовано повний збіг першої і третьої лексеми.
4. Якщо так, то чи є другим компонентом одне із зафіксованих у реченнях цього типу слів: *є, це, то, не, як, так*. Якщо ця умова виконана, то речення вважаємо синтаксичним фразеологізмом і переходимо до наступних трьох слів у реченні. Якщо у реченні більше немає слів, то беремо наступне речення і починаємо цикл перевірки спочатку.
5. Якщо умову щодо другого компонента не виконано, перевіряємо чи є перше і третє слово словоформами одного слова. В УНЛК це можна перевірити за збігом квазіоснов. Якщо так, то визначаємо, чи є друге слово прийменником *над, з/із/зі, між, серед.* За виконання цієї умови кваліфікуємо аналізовану послідовність трьох слів як синтаксичний фразеологізм.
6. Якщо ні, то зіставляємо перше і друге слово і встановлюємо, чи належать вони до форм одного слова. Якщо умову дотримано, з'ясовуємо, чи є третє слово сполучником *а, але, зате, проте, однак.* Якщо так, то вважаємо, що це синтаксичний фразеологізм.
7. Далі переходимо до аналізу речень без повторів. Передусім перевіряємо, чи зафіксоване поєднання першого та другого слова в укладеному нами реєстрі двочленних стрижневих компонентів синтаксичних фразеологізмів (див. табл. 1). Якщо так, то маємо синтаксичний фразеологізм.
8. Якщо такого поєднання в реєстрі поки що не зафіксовано, встановлюємо частиномовну належність першого та другого слів. В УНЛК це можна зробити за допомогою квазіфлексій. Якщо ці слова належать до службових частин мови, займенників або прислівників, то є потреба в наступному кроці алгоритму.
9. Для незафіксованого в реєстрі поєднання слів виконуємо таку послідовність дій: визначаємо в корпусі текстів абсолютну частоту першого слова *f(x)*, абсолютну частоту другого слова *f(y)* та абсолютну частоту конструкції загалом *f(x, y)*, підставляємо отримані числа до формули (1) та обраховуємо коефіцієнт *MI.*
10. Якщо отриманий результат MI ≥9, аналізовану конструкцію кваліфікуємо як синтаксичний фразеологізм і переходимо до аналізу наступних трьох слів або наступного речення за їхньої наявності.

На виході алгоритму маємо файл з усіма контекстами вживання синтаксичних фразеологізмів, отриманими на підставі аналізу корпусу текстів.

Схема 1

Алгоритм розпізнавання синтаксичних фразеологізмів (із одно- або двочленним стрижневим компонентом) у корпусі текстів



Отже, запропонований алгоритм розпізнавання синтаксичних фразеологізмів об'єднує дві гілки, побудовані відповідно для речень із одночленним та речень із двочленним стрижневим компонентом. Перша гілка ґрунтується на наявності повторів слів (збігу слів повністю або наявність словоформ одного слова) та визначеному на попередніх етапах дослідження переліку стрижневих компонентів (*є, це, то, не, так; як; з/із/зі, між, над, серед; а, але, зате, однак, проте*).

Друга гілка створена для іншого типу синтаксичних фразеологізмів – із двочленним стрижневим компонентом. Вона враховує такі властивості аналізованих одиниць: наявність поєднань службових частин мови, службової частини мови із займенником або прислівником, займенника та прислівника; відповідність поєднань слів укладеному реєстру стрижневих компонентів синтаксичних фразеологізмів, показник асоціації *mutual information* ≥ 9.

На подальшому етапі дослідження плануємо створити робочу версію програми автоматичного розпізнавання синтаксичних фразеологізмів на основі УНЛК та здійснити корекцію алгоритму відповідно до отриманих результатів роботи програми.

References

Church, Kenneth Ward, and Patrick Hanks. "Word Association Norms, Mutual Information, and Lexicography". *Computational Linguistics* 16(1) (1990): 22–29. Print.

Gusev, Vladimir, and Natal'ja Salomatina. "Algoritm Vyjavlenija Ustojchivyh Slovosochetanij s Uchetom Ih Variativnosti (Morfologicheskoj i Kombinatornoj) (Algorithm of Identification of Fixed Phrases Considering Their Variability (Morphological and Combinatorial))." *Trudy Mezhdunar. Konf. Dialog-2004, 2–7 Ijunja 2004, Verhnevolzhskij* (*Proceedings of International Conference Dialog-2004, June 2–7, 2004, Verhnevolzhskij*). Moscow: Nauka, 2004. 530–535. Print.

GOST 19.701-90. *Skhemy Algoritmov, Programm, Dannykh i Sistem. Uslovnye Oboznachenija i Pravila Vypolnenija* (State Standard 19.701-90. *Schemes of Algorithms, Programs, Data and Systems. Conventional Signs and Rules of Performance*). Web. 5 Oct. 2016.

Lychuk, Mariya. *Stupeni Frazeolohizatsiyi Rechen'* (*Stages of Sentence Phraseologization*). Diss. Chernivtsi National U named after Yuriy Fedkovych, 2001. Chernivtsi, 2001. Abstract. Print.

Lychuk, Mariya, and Vasyl Shynkaruk. *Stupeni Frazeolohizatsiyi Rechen'* (*Stages of Sentence Phraseologization*). Chernivtsi: Ruta, 2001. Print.

Russkaya Grammatika : V 2 Tt. (*Russion Grammar* : In 2 Vol.). Vol. 2. Moscow: Nauka, 1980. Print.

Shmelev, Dmitriy. *Sintaksicheskaya Chlenimost' Vyskazyvaniya v Sovremennom Russkom Yazyke* (*Syntactic Divisibility of Utterance in the Modern Russian Language*). Moscow: KomKniga, 2006. Print.

Shvedova, Nataliya. *Ocherki po Sintaksisu Russkoy Razgovornoy Rechi* (*Esssays about Syntax of Russian Colloquial Speech*). Moscow: Izd-vo AN SSSR, 1960. Print.

Sytar, Hanna. "Status Syntaksychnykh Frazeolohizmiv u Systemi Frazeolohichnykh Odynyts' (The Status of Syntactic Idioms in the System of Phraseological Units)." *Visnyk Donets'koho Natsional'noho Universytetu. Seriya B. Humanitarni Nauky* (*The Bulletin of Donetsk National University. Series B. Humanities*) 2 (2011): 66–74. Print.

Sytar, Hanna. "Strukturni y Semantychni Typy Syntaksychnykh Frazeolohizmiv v Ukrayins'kiy Movi (Structural and Semantic Types of Syntactic Idioms in the Ukrainian Language)." *Movoznavchyy Visnyk* (*Linguistic Bulletin*)12-13 (2011): 178–181. Print.

Sytar, Hanna. "Statystychni Kryteriyi Analizu Syntaksychnykh Frazeolohizmiv (Statistical Criteria of Analysis of Syntactic Idioms)." *Visnyk Donets'koho Natsional'noho Universytetu. Seriya B. Humanitarni Nauky* (*The Bulletin of Donetsk National University. Series B. Humanities*) 1-2 (2015): 245–256. Print.

Sytar, Hanna. "Syntaksychni Frazeolohizmy i Pretsedentni Fenomeny: Zony Peretynu (Syntactic Idioms and Precedent Phenomena: Intersection Zones)." *Linhvistychni Studiyi / Linguistic Studies* 31 (2016): 20–25. Print.

Sitar, Anna. "Modeli Sintaksicheskih Frazeologizmov s Povtorami v Ukrainskom Jazyke: Popytka Klassifikacii (Models of Syntactic Idioms with Repeats in the Ukrainian Language: Attempt of Classification)." *Jazyk, Soznanie, Kommunikacija* (*Language, Consciousness, Communication*) 47 (2013): 485–493. Print.

Velichko, Alla. *Sintaksicheskaya Frazeologiya Dlya Russkikh i Inostrantsev* (*Syntactic Phraseology for Russions and Foreigners*). Moscow: Izd-vo MGU, 1996. Print.

Vsevolodova, Maya, and Yon Lim Su. *Printsipy Lingvisticheskogo Opisaniya Sintaksicheskikh Frazeologizmov: Na Materiale Sintaksicheskikh Frazeologizmov so Znacheniem Otsenki* (*The Principles of Linguistic Description of Syntactic Idioms : Based on the Syntactic Idioms with Evaluative Meaning*). Moscow: Maks PRESS, 2002. Print.

Надійшла до редакції 29 вересня 2016 року.

ALGORITHM OF SYNTACTIC IDIOMS RECOGNITION IN THE CORPUS OF TEXTS: ATTEMPT OF CREATION

Hanna Sytar

Department of General and Applied Linguistics and Slavonic Philology, Vasyl' Stus Donetsk National University, Vinnytsia, Ukraine

Abstract

Background: Attention of national and foreign researchers was focused so far on structural and semantic features of syntactic idioms. Automatic analysis of these peculiar units that are on the verge of syntax and phraseology still was not carried out in the scientific literature. This issue requires a theoretical understanding and practical implementation.

Purpose: To create an algorithm of recognition of syntactic idioms with one- or two-term core component in the corpus of texts.

Results: Based on the results of previous theoretical studies we highlighted a number of formal and statistical criteria that enable to distinguish syntactic idioms from other language units in the corpus of Ukrainian-language texts.

The author developed a block diagram of syntactic idioms recognition, incorporating two branches constructed accordingly for the sentences with one-term and sentences with two-term core component. The first branch is based on the presence of word repeats (full words concurrence or presence of other word forms of the word) and the list of core components determined on previous stages of the study (*є, це, то, не, так; як; з/із/зі, між, над, серед; а, але, зате, однак, проте*).

The second branch was created for another type of syntactic idioms – one with a two-term core component. It takes into account the following properties of the analyzed units: the presence of combinations of service parts of speech, service parts of speech with pronoun or adverb, pronoun and adverb; compliance of words combinations with the register of the syntactic idioms core components currently comprising 92 structures; association measure of mutual information ≥ 9, etc.

Discussion: Offered algorithm enables automatic identification of syntactic idioms in the corpus of texts and removal of contexts of their use, it can be used to improve the procedure of automatic text processing and creation of automated translation systems.

In the future phase of the study we plan to develop the release version of program of syntactic idioms automatic recognition based on Ukrainian National Linguistic Corpus and to correct the algorithm according to the achieved results of the program work.

Keywords: algorithm, corpus of texts, model of sentence, syntactic idiom, sentence with phraseological structure, the Ukrainian language.

Vitae

Hanna Sytar is PhD of Philology, Assosiate Professor, Doctoral Candidate of Department of General and Applied Linguistics and Slavonic Philology at Vasyl' Stus Donetsk National University. Her areas of research interests include syntax, semantics, pragmatics, construction grammar, applied linguistics.

Correspondence: h.v.sytar@donnu.edu.ua