является спорным. Ошибки в результатах кластеризации могут быть устранены экспертом или аналитиком.

Выводы. Кластеризация делается для того, чтобы сократить объем входных данных и в случае возникновения спорных элементов может вызывать потерю соответствующих им документов, что критично при работе с малыми коллекциями. Первостепенный этап выборки начальных центроидов может быть реализован как функцией рандомного распределения, так и по определенному алгоритму, в зависимости от поставленных целей. Конечный результат кластеризации (количество итераций, количество создаваемых кластеров и правильность отнесения к ним обрабатываемых данных), в равной степени зависит от правил начального выбора центроидов и правил остановки процесса кластеризации.

Список использованных источников

1. *Словари* и энциклопедии. Кластерный анализ [Электронный ресурс]. – Режим доступа : http://dic.academic.ru/dic.nsf/enc_psychology/349/Кластерный.

2. *Задачи* кластерного анализа [Электронный ресурс]. – Режим доступа : http://ru.science.wikia.com/wiki/Кластерный_анализ.

3. *Методы* кластерного анализа [Электронный ресурс]. – Режим доступа : http://bug.kpi.ua/stud/work/RGR/DATAMINING/clusteranalisismethods.html.

4. *Факторный*, дискриминантный и кластерный анализ : пер. с англ. / Дж.-О. Ким, Ч. У. Мьюллер, У. Р. Клекка и др. ; под ред. И. С. Енюкова. – М. : Финансы и статистика, 1989. – 215 с.

5. *SVM-Light* Support Vector Machine [Электронный ресурс]. – Режим доступа : http://www.svmlight.joachims.org.

6. *Литвинов В. В.* Автоматизованная система обработки динамических коллекций разноязычных текстовых документов по морскому и речному делу / В. В. Литвинов, О. П. Мойсеенко // Математические машины и системы. – 2014. – № 2. – С. 59–64.

УДК 004.519.7(045)

А.І. Вавіленкова, канд. техн. наук
Національний авіаційний університет, м. Київ, Україна

АНАЛІЗ МЕТОДІВ ПОШУКУ СИНОНІМІВ В ЕЛЕКТРОННИХ ДОКУМЕНТАХ

А.И. Вавиленкова, канд. техн. наук
Национальный авиационный университет, г. Киев, Украина

АНАЛИЗ МЕТОДОВ ПОИСКА СИНОНИМОВ В ЭЛЕКТРОННЫХ ДОКУМЕНТАХ

Anastasiia Vavilenkova, PhD in Technical Sciences
National Aviation University, Kyiv, Ukraine

ANALYSIS OF THE METHODS OF SEARCHING SYNONYMS IN ELECTRONIC DOCUMENTS

Проаналізовано статистичні засоби пошуку синонімів у природномовних текстах, розглянуто алгоритми пошуку синонімів, що ґрунтуються на використанні тезауруса мови. Запропоновано формальні умови виявлення синонімічних конструкцій через використання логіко-лінгвістичних моделей. Це стало можливим завдяки дослідженню трансформацій слів. Проведено статистичний аналіз використання синонімічних конструкцій у різних типах текстових документів.

Ключові слова: *природна мова, синоніми, конверсиви, трансформація, аналіз тексту, логіко-лінгвістичні моделі, електронні документи.*

Осуществлен анализ статистических методов поиска синонимов в текстах естественного языка, рассмотрены алгоритмы поиска синонимов, которые основываются на использовании тезауруса речи. Предложены формальные условия нахождения синонимических конструкций путем использования логико-лингвистических моделей. Это стало возможным благодаря исследованию трансформаций слов. Проведен статистический анализ использования синонимических конструкций в различных типах текстовых документов.

Ключевые слова: *естественный язык, синонимы, конверсивы, трансформация, анализ текста, логико-лингвистическая модель, электронные документы.*

The article presents the analysis of the statistical methods of searching for synonyms in natural language text .The algorithms are based on the use of a thesaurus of speech. The proposed formal conditions for searching synonymous

constructions are possible by means of logic-linguistic models. All that can be done due to the study of transformations of words. Article provides with statistical analysis of the use of synonymous structures in different types of text documents.

Key words: *natural language, synonyms, conversions, transformation, text analysis, logic-linguistic models, electronic documents.*

Постановка проблеми. Проблема пошуку синонімічних конструкцій займає одну з першочергових позицій у процесі здійснення змістовної обробки електронних документів. Адже методи пошуку взаємозамінних синтаксичних складових, що використовуються сьогодні пошуковими системами, – статистичні і не враховують зміст текстової інформації. Через це інформаційний простір наповнений величезною кількістю електронних документів, які дублюються повністю, дещо змінені чи трансформовані за рахунок неточного перекладу, а на запит користувача видаються відповіді з все меншою релевантністю.

Для змістовного автоматичного аналізу електронних документів необхідно розробити такий формальний апарат, який об'єднував би в собі всі можливі способи подання контексту (з позиції лінгвістики) та математичні методи їх виявлення. Для знаходження текстових збігів та логічних суперечностей потрібні алгоритми ідентифікації синонімічних конструкцій, що є основою порівняльного аналізу за змістом.

Синонімами вважають вирази, які збігаються або близькі за лексичним значенням, здатні замінити одне одного в деяких контекстах [1]. Якщо формалізувати умови виявлення синонімів у природномовних текстах, то стає можливою автоматизація порівняльного аналізу електронних документів за змістом.

Аналіз останніх досліджень і публікацій. Аналіз досліджень у сфері комп'ютерної лінгвістики показав, що чисельні теорії та експерименти у сфері аналітичної обробки текстової інформації досі не дали можливість створити автоматизовану систему змістовного аналізу текстових документів. На заваді стають такі проблеми, як знаходження синонімів, автоматичне зняття омонімії, інверсний порядок слів у реченні, логічні суперечності, авторські знаки у текстах та ін.

Зокрема, вирішенням перерахованих вище проблем займаються Н.Ф. Алєфіренко, намагаючись через різні інтерпретації синтаксичного значення дійти до семантичної суті [2], Dirk Geeraerts у роботі «Cognitive linguistics: basic readings research» [3], М.В. Нікітін, досліджуючи компоненти змістовної структури поняття та ієрархії узагальнень [4].

М.О. Кронгауз [5] у своїх дослідженнях позиціонує синоніми як слова, що повинні відноситися до тієї ж самої частини мови, спираючись на роботи Ю.Д. Апресяна [6], наполягає на однаковій кількості активних семантичних валентностей та співвіднесенні однакових валентностей з однаковими ролями.

Американські лінгвісти Dan Jurafsky та Christopher Manning [7] пропонують курс лекцій з обробки природномовних текстів, у якому узагальнюють всю відому на сьогодні інформацію щодо морфологічного, синтаксичного, семантичного та когнітивного аналізу текстової інформації.

Виділення не вирішених раніше частин загальної проблеми. Саме відсутність формальних засобів виявлення синонімічних конструкцій є основною проблемою на шляху автоматизованого порівняння текстових документів за змістом. Тому матеріали дослідження спрямовані на розроблення формального апарату, який дозволить знаходити в електронних документах синонімічні конструкції, враховуючи всі можливі їх вираження.

Мета статті. Метою роботи є аналіз методів пошуку синонімів в електронних документах, виявлення недоліків наявних методів та розроблення нових алгоритмів, що забезпечили б можливість виявлення змісту текстової інформації та дали б можливість автоматично порівнювати контекст текстових документів.

Виклад основного матеріалу. Сучасні системи обробки текстової інформації вирішують проблему виявлення синонімічних конструкцій як завдання пошуку фрагментів тексту, що збігається з шаблоном. Зокрема, для цього використовуються алгоритми

пошуку підрядочка в рядочку, наприклад, алгоритм Карпа-Рабіна, метод шинглів, алгоритм Кнута-Морріса-Пратта, алгоритм Бойєра-Мура [8], методи лексичних сигнатур, алгоритм виявлення інформаційних сюжетів [9].

У комп'ютерній лінгвістиці застосовується поняття відстані Левенштейна, яке означає мінімальну кількість операцій вставки, видалення та заміни одного символу на інший, що необхідно для перетворення одного рядочка в інший [10].

Нехай T_1 і T_2 два рядочки довжиною l_1 та l_2 відповідно, тоді відстань Левенштейна $d(T_1,T_2)$ розраховують за формулою $d(T_1,T_2)=D(l_1,l_2)$, де

$$D(l_1,l_2)=\begin{cases}0,\ i=0,\ j=0,\\ i,\ \ j=0, i>0,\\ j,\ i=0,\ j>0,\\ \min(D(i,\ j-1)+1, D(i-1,\ j)+1, D(i-1,\ j-1)+m(T_1[i],T_2[\,j\,])),\ j>0, i>0;\end{cases}$$

i – номер символу першого рядочка T_1;

j – номер символу другого рядочка T_2;

$m(T_1[i],T_2[\,j\,])=0$, якщо $T_1[i]=T_2[\,j\,]$;

$m(T_1[i],T_2[\,j\,])=1$, якщо $T_1[i]\neq T_2[\,j\,]$;

крок по i означає видалення символу з першого рядочка;

крок по j означає вставку символу в перший рядок;

крок по обох символах означає заміну символу або відсутність змін.

Якщо переставити слова місцями, то відстань Левенштейна між ними буде великою, що говоритиме про їх несинонімічність. Також відстань між схожими довгими словами буде більшою, ніж між абсолютно не схожими короткими. Це свідчить про те, що аналіз слів у рядочках відбувається суто статистично, не враховуючи зміст слів природної мови.

Учені у сфері комп'ютерної лінгвістики [7] розрізняють два класи алгоритмів пошуку синонімів.

1. *Алгоритми, що використовують для пошуку тезауруса природної мови.*

а) Два концепти вважаються однаковими, якщо вони розміщені один біля одного в ієрархії тезауруса, тобто мають між собою коротку відстань, а довжина шляху концепту до самого себе дорівнює 1. Наприклад, нехай у тезаурусі української мови існує ієрархія синонімів до слова «говорити» (рис. 1).

Рис. 1. Ієрархія синонімів до слова «говорити»

Якщо позначити два концепти як c_1 та c_2, то відстань між ними в тезаурусі розраховується як:

$$pathlen(c_1, c_2) = 1 + q,$$

де q – кількість позицій найкоротшого шляху на орієнтованому графі від концепту c_1 до c_2.

Тоді синонімічна відстань у тезаурусі між концептами c_1 та c_2 дорівнює:

$$simpath(c_1, c_2) = \frac{1}{pathlen(c_1, c_2)}.$$

Наприклад, для ієрархії синонімів, зображеній на рис. 1, синонімічна відстань між словами буде розрахована таким чином:

simpath (розповідати, казати) = 1/2=0,5.

simpath (розповідати, повідати) = 1/3=0,33.

simpath (розповідати, гомоніти) = 1/5=0,2.

simpath (розповідати, патякати) = 1/8=0,125.

Розрахована таким чином синонімічна відстань носить абстрактний характер, оскільки кожен текст носить свої відтінки використання тих чи інших однозначних слів, які не можуть бути враховані при створенні ієрархії синонімів у тезаурусі. До того ж створення таких ієрархій – дуже трудомісткий процес, що потребує зусиль багатьох лінгвістів.

б) Слова вважаються однаковими за змістом, якщо мають однакові визначення у тезаурусі, а синонімічність визначається за допомогою підрахунків.

Для кожного концепту в ієрархії підраховується кількість можливих визначень по горизонталі:

$$fq(c) = \frac{\sum_{w \in words(c)} count(w)}{N},$$

де w – слово в ієрархії синонімів;

$words(c)$ – набір усіх слів, що вважаються дочірніми від концепту c;

N – загальна кількість слів в ієрархії.

Наприклад, для ієрархії (рис. 1) можна побудувати набори слів:

words(«базікати»)={жебоніти патякати},

words(«розмовлятиі»)={розповідати, свідчити, повідати, казати, мовити}.

2. *Розподілені алгоритми* (визначають, чи мають слова однаковий розподілений контекст).

Для розподілених алгоритмів характерно те, що в основі ідентичності двох слів лежить загальна інформація, тобто чим більше слова мають спільної загальної інформації, тим більше схожі вони за змістом. Розрахунок синонімічності двох концептів можна здійснювати за такими формулами:

а) за методом Різника:

$$sim_{reznik}(c_1, c_2) = -\log fq(lcs(c_1, c_2)),$$

б) за методом Деканга Ліна враховується не тільки схожість, але й те, що чим більше різного в загальній інформації про концепти, тим менш вони схожі, а схожість розраховується як відношення кількості інформації, необхідної для опису загального, до інформації, потрібної для повного опису того, чим є концепти:

$$sim_{Lin}(c_1, c_2) = \frac{2 \log fq(lcs(c_1, c_2))}{\log fq(c_1) + fqP(c_2)},$$

де $lcs(c_1, c_2)$ – найнижчий опис в ієрархії, який описує і концепт c_1, і концепт c_2.

Усі описані вище методи пошуку синонімів дієві з тією чи іншою ймовірністю лише для слів і не можуть бути застосовані до речень природної мови.

Формальні умови виявлення синонімічних конструкцій завдяки використанню логіко-лінгвістичних моделей. Нехай розглядаються два речення (рядочка) T_1 і T_2:

$T_1[i]$ – символ першого рядочка, $i=\overline{1,l_1}$;

$T_2[j]$ – символ другого рядочка, $j=\overline{1,l_2}$;

S_k – слово першого рядочка, $k=\overline{1,n}$, n – загальна кількість слів у рядочку T_1;

S_r – слово другого рядочка, $r=\overline{1,m}$, m – загальна кількість слів у рядочку T_2.

Кожному слову речення природної мови S ставиться у відповідність набір характеристик [11]:

$$Z(S)=\{cm,g,n,k2,t,h,l,ch\},$$

де $cm=\overline{1,11}$ – граматична характеристика, що позначає частину мови, кожному цифровому значенню якої відповідає іменник, прикметник, числівник, займенник, дієслово, дієприкметник, дієприслівник, прислівник, прийменник, сполучник або частка відповідно;

$g=\overline{1,7}$ – морфологічна ознака, що відповідає за відмінок;

$n=\overline{1,2}$ – граматичний параметр, що означає число;

$k2=\overline{1,4}$ – граматичний параметр, що означає рід;

$t=\overline{0,3}$ – граматичний параметр, що означає час;

$h=\overline{1,3}$ – граматичний параметр, що означає спосіб;

$l=\overline{1,3}$ – граматичний параметр, що означає особу;

$ch=\overline{1,5}$ – параметр, що означає синтаксичну роль (підмет, присудок, додаток, означення, обставина).

Можна сформулювати умови, за яких речення природної мови будуть формально вважатися однаковими за змістом. Для знаходження синонімічних конструкцій не має значення порядок розгляду речень, тобто умови можна застосовувати й у зворотну сторону, коли першим буде вважатися речення T_2, а другим T_1.

1. Умова, пов'язана з існуванням у флективних мовах такого явища, як синтаксична деривація, ще називається транспозиційною деривацією: перехід слова з однієї частини мови в іншу без зміни його лексичного значення.

а) Якщо в першому реченні T_1 знайдено послідовність слів $S_k \to S_{k+1} \to S_{k+2}$, що володіють характеристиками

$$Z_k(S_k)=\{cm_k,1,n_k,k2_k,0,0,l_k,1\},$$
$$Z_{k+1}(S_{k+1})=\{5,0,n_k,k2_k,t_{k+1},h_{k+1},l_k,2\},$$
$$Z_{k+2}(S_{k+2})=\{cm_{k+2},g_{k+2}\neq 1,n_{k+2},k2_{k+2},t_{k+2},h_{k+2},l_{k+2},ch_{k+2}\},$$

і такій послідовності можна поставити у відповідність послідовність слів $S_r \to S_{r+1} \to S_{r+2} \to S_{r+3}$ другого речення T_2, таку, що $S_r \equiv S_k$, $S_{r+1} \neq S_{k+1}$, $\widehat{S}_{r+2} \equiv \widehat{S}_{k+1}$ (де $\widehat{S}_{r+2}, \widehat{S}_{k+1}$ – спільнокореневі слова), $S_{r+3} \equiv S_{k+2}$ з характеристиками

$$Z_r(S_r)=\{cm_k,1,n_k,k2_k,0,0,l_k,1\},$$
$$Z_{r+1}(S_{r+1})=\{5,0,n_k,k2_k,t_{k+1},h_{k+1},l_k,2\},$$
$$Z_{r+2}(S_{r+2})=\{1,g_{r+2}\neq 1,n_{r+2},k2_{r+2},0,0,l_{r+2},2\},$$
$$Z_{r+3}(S_{r+3})=\{cm_{k+2},g_{k+2}\neq 1,n_{k+2},k2_{k+2},t_{k+2},h_{k+2},l_{k+2},ch_{k+2}\}$$

відповідно, то речення T_1 та T_2 тотожні за змістом.

Наприклад, «*Традиція розпочинається у минулому столітті*» та «*Традиція бере початок у минулому столітті*».

Логіко-лінгвістичні моделі речень мають вигляд:

Розпочинається (традиція, століття {минулому}), $P_1(x_1, x_2\{c_{21}\})$,

Бере& початок (традиція, століття {минулому}), $P_2 \,\&\, \widehat{P}_1(x_1, x_2\{c_{21}\})$.

У цьому разі суб'єкти та об'єкти логіко-лінгвістичних моделей однакові, а предикати відповідають схемі: $P_1 \equiv P_2 \,\&\, \widehat{P}_1$, де $\widehat{P}_1$ – слово, спільнокореневе до P_1.

б) Якщо в першому реченні T_1 знайдено послідовність слів $S_k \to S_{k+1} \to S_{k+2}$, що володіють характеристиками

$$Z_k(S_k) = \{cm_k, 1, n_k, k2_k, 0, 0, l_k, 1\},$$
$$Z_{k+1}(S_{k+1}) = \{5, 0, n_k, k2_k, t_{k+1}, h_{k+1}, l_k, 2\},$$
$$Z_{k+2}(S_{k+2}) = \{cm_{k+2}, g_{k+2} \neq 1, n_{k+2}, k2_{k+2}, 0, 0, l_{k+2}, 3\},$$

і такій послідовності можна поставити у відповідність послідовність слів $S_r \to S_{r+1} \to S_{r+2}$ другого речення T_2, таку, що $S_r \equiv S_k$, $\widehat{S}_{r+1} \equiv \widehat{S}_{k+1}$, $S_{r+2} \equiv S_{k+2}$ з характеристиками

$$Z_r(S_r) = \{cm_k, 1, n_k, k2_k, 0, 0, l_k, 1\},$$
$$Z_{r+1}(S_{r+1}) = \{6, 1, n_k, k2_k, 0, 0, l_k, 2\},$$
$$Z_{r+2}(S_{r+2}) = \{cm_{k+2}, g_{r+2} \neq g_{k+2}, n_{k+2}, k2_{k+2}, 0, 0, l_{k+2}, 3\}$$

відповідно, то речення T_1 та T_2 тотожні за змістом.

Наприклад, речення «*Робота цікавить викладача*» і «*Робота цікава викладачу*». Логіко-лінгвістичні моделі речень мають вигляд:

Цікавить (робота, викладача), $P_1(x_1, x_2)$,

Цікава (робота, викладачу), $\widehat{P}_1(x_1, x_2)$.

Аналогічно першій умові, логіко-лінгвістичні моделі речень однакові, а предикати відповідають схемі: $P_1 \equiv \widehat{P}_1$.

в) Якщо в першому реченні T_1 знайдено послідовність слів $S_k \to S_{k+1} \to S_{k+2}$, що володіють характеристиками

$$Z_k(S_k) = \{cm_k, 1, n_k, k2_k, 0, 0, l_k, 1\},$$
$$Z_{k+1}(S_{k+1}) = \{5, 0, n_k, k2_k, t_{k+1}, h_{k+1}, l_k, 2\},$$
$$Z_{k+2}(S_{k+2}) = \{cm_{k+2}, g_{k+2} \neq 1, n_{k+2}, k2_{k+2}, 0, 0, l_{k+2}, 3\},$$

і такій послідовності можна поставити у відповідність послідовність слів $S_r \to S_{r+1} \to S_{r+2} \to S_{r+3} \to S_{r+4}$ другого речення T_2, таку, що $S_{r+1} \equiv S_{k+2}$, $\widehat{S}_{r+3} \equiv \widehat{S}_{k+2}$, $S_{r+4} \equiv S_k$ з характеристиками

$$Z_r(S_r) = \{9, g_{r+1}, 0, 0, 0, 0, 0, 0\},$$
$$Z_{r+1}(S_{r+1}) = \{cm_{k+2}, g_{k+2} \neq 1, n_{k+2}, k2_{k+2}, 0, 0, l_{k+2}, 3\},$$
$$Z_{r+2}(S_{r+2}) = \{cm_{r+2}, 1, n_{r+2}, k2_{r+2}, 0, 0, l_{r+2}, 2\},$$
$$Z_{r+3}(S_{r+3}) = \{9, g_{r+4}, 0, 0, 0, 0, 0, 0\},$$
$$Z_{r+4}(S_{r+4}) = \{cm_k, g_{r+4} \neq 1, n_k, k2_k, 0, 0, l_k, 3\}$$

відповідно, то речення T_1 та T_2 тотожні за змістом.

Наприклад, речення «*Робота цікавить викладача*» і «*У викладача цікавість до роботи*». Логіко-лінгвістичні моделі речень мають вигляд:

Цікавить (робота, викладача), $P_1(x_1, x_2)$,

Цікавість (, викладача, роботи), $\widehat{P}_1(,x_2,x_1)$.

Таким чином, предикатна змінна (суб'єкт) у другому реченні відсутня, а предикати відповідають схемі: $P_1 \equiv \widehat{P}_1$, де $\widehat{P}_1$ – слово, спільнокореневе до P_1.

2. Синонімічними за змістом можна вважати речення, у яких змінені синтаксичні позиції лексичних морфем, а денотативне значення залишається незмінним, з рахуванням таких умов.

а) Якщо в першому реченні T_1 знайдено послідовність слів $S_k \to S_{k+1} \to S_{k+2}$, що володіють характеристиками

$$Z_k(S_k) = \{cm_k, 1, n_k, k2_k, 0, 0, l_k, 1\},$$
$$Z_{k+1}(S_{k+1}) = \{5, 0, n_k, k2_k, t_{k+1}, h_{k+1}, l_k, 2\},$$
$$Z_{k+2}(S_{k+2}) = \{cm_{k+2}, g_{k+2} \neq 1, n_{k+2}, k2_{k+2}, 0, 0, l_{k+2}, 3\},$$

і такій послідовності можна поставити у відповідність послідовність слів $S_r \to S_{r+1} \to S_{r+2}$ другого речення T_2, таку, що $S_r \equiv S_{k+2}$, $S_{r+1} \equiv S_{k+1}$, $S_{r+2} \equiv S_k$ з характеристиками

$$Z_r(S_r) = \{cm_{k+2}, 1, n_{k+2}, k2_{k+2}, 0, 0, l_{k+2}, 1\},$$
$$Z_{r+1}(S_{r+1}) = \{5, 0, n_r, k2_r, t_{r+1}, h_{r+1}, l_r, 2\},$$
$$Z_{r+2}(S_{r+2}) = \{cm_k, g_{r+1} \neq 1, n_k, k2_k, 0, 0, l_k, 3\}$$

відповідно, то речення T_1 та T_2 тотожні за змістом.

Наприклад, речення «*Експерти використовують методи*» та «*Методи використовуються експертами*». Логіко-лінгвістичні моделі речень мають вигляд:

Використовують (експерти, методи), $P_1(x_1, x_2)$,

Використовуються (методи, експертами), $P_1(x_2, x_1)$.

б) Якщо в першому реченні T_1 знайдено послідовність слів $S_k \to S_{k+1} \to S_{k+2}$, що володіють характеристиками

$$Z_k(S_k) = \{cm_k, 1, n_k, k2_k, 0, 0, l_k, 1\},$$
$$Z_{k+1}(S_{k+1}) = \{5, 0, n_k, k2_k, t_{k+1}, h_{k+1}, l_k, 2\},$$
$$Z_{k+2}(S_{k+2}) = \{cm_{k+2}, g_{k+2} \neq 1, n_{k+2}, k2_{k+2}, 0, 0, l_{k+2}, 3\},$$

і такій послідовності можна поставити у відповідність послідовність слів $S_r \to S_{r+1} \to S_{r+2} \to S_{r+3}$ другого речення T_2, таку, що $S_r \equiv S_{k+2}$, $S_{r+1} \equiv S_{k+1}$, $S_{r+3} \equiv S_k$ з характеристиками

$$Z_r(S_r) = \{cm_{k+2}, 1, n_{k+2}, k2_{k+2}, 0, 0, l_{k+2}, 1\},$$
$$Z_{r+1}(S_{r+1}) = \{5, 0, n_r, k2_r, t_{r+1}, h_{r+1}, l_r, 2\},$$
$$Z_{r+2}(S_{r+2}) = \{9, g_{r+3}, 0, 0, 0, 0, 0, 0\},$$
$$Z_{r+3}(S_{r+3}) = \{cm_k, g_{r+3} \neq 1, n_k, k2_k, 0, 0, l_k, 3\}$$

відповідно, то речення T_1 та T_2 тотожні за змістом.

Наприклад, речення «*Магніт притягує залізо*» та «*Залізо притягується до магніту*». Логіко-лінгвістичні моделі речень мають вигляд:

Притягує (магніт, залізо), $P_1(x_1, x_2)$,

Притягується (залізо, магніту), $P_1(x_2, x_1)$.

Аналіз логіко-лінгвістичних моделей у двох описаних вище ситуаціях дає змогу зробити висновок про те, що якщо предикатна змінна (суб'єкт) одного речення дорів-

нює предикатній змінній (аргументу) другого речення, і навпаки, а також рівні предикати, то такі речення тотожні за змістом.

в) Якщо в першому реченні T_1 знайдено послідовність слів $S_k \to S_{k+1} \to S_{k+2} \to S_{k+3}$, що володіють характеристиками

$$Z_k(S_k) = \{cm_k, 1, n_k, k2_k, 0, 0, l_k, 1\},$$
$$Z_{k+1}(S_{k+1}) = \{5, 0, n_k, k2_k, t_{k+1}, h_{k+1}, l_k, 2\},$$
$$Z_{k+2}(S_{k+2}) = \{cm_{k+2}, g_{k+2} \neq 1, n_{k+2}, k2_{k+2}, 0, 0, l_{k+2}, 3\},$$
$$Z_{k+3}(S_{k+3}) = \{cm_{k+3}, g_{k+3} \neq 1, n_{k+3}, k2_{k+3}, 0, 0, l_{k+3}, 3\},$$

і такій послідовності можна поставити у відповідність послідовність слів $S_r \to S_{r+1} \to S_{r+2} \to S_{r+3} \to S_{r+4}$ другого речення T_2, таку, що $S_r \equiv S_k$, $S_{r+1} \equiv S_{k+1}$, $S_{r+2} \equiv S_{k+3}, S_{r+4} \equiv S_{k+2}$ з характеристиками

$$Z_r(S_r) = \{cm_k, 1, n_k, k2_k, 0, 0, l_k, 1\},$$
$$Z_{r+1}(S_{r+1}) = \{5, 0, n_k, k2_k, t_{k+1}, h_{k+1}, l_k, 2\},$$
$$Z_{r+2}(S_{r+2}) = \{cm_{k+3}, (g_{r+2} \neq 1) \& (g_{r+2} \neq g_{k+3}), n_{k+3}, k2_{k+3}, 0, 0, l_{k+3}, 3\},$$
$$Z_{r+3}(S_{r+3}) = \{9, g_{r+4}, 0, 0, 0, 0, 0, 0\},$$
$$Z_{k+4}(S_{k+4}) = \{cm_{k+2}, (g_{r+4} \neq 1) \& (g_{r+4} \neq g_{k+2}), n_{k+2}, k2_{k+2}, 0, 0, l_{k+2}, 3\}$$

відповідно, то речення T_1 та T_2 тотожні за змістом.

Такий запис показує, що лівий та правий трансформи описують ситуацію того самого класу. В результаті використання таких синонімічних конструкцій зміщується логічний акцент, змінюються синтаксичні позиції лексичних морфем, проте зберігається зміст.

Наприклад, речення «*Учені постачають бібліотеки книжками*» та «*Учені постачають книжки до бібліотек*». Логіко-лінгвістичні моделі речень мають вигляд:

Постачають (вчені, бібліотеки, книжками), $P_1(x_1, x_2, x_3)$,

Постачають (вчені, книжки, бібліотек), $P_1(x_1, x_3, x_2)$.

Такий вид синонімії відображається в логіко-лінгвістичних моделях через заміну предикатних змінних (аргументів).

3. Трансформації, в яких змінюється прийменниково-відмінкова форма не більше ніж у одного іменника, вважаються синонімічними.

Якщо в першому реченні T_1 знайдено послідовність слів $S_k \to S_{k+1}$, що володіють характеристиками

$$Z_k(S_k) = \{5, 0, n_k, k2_k, t_k, h_k, l_k, 2\},$$
$$Z_{k+1}(S_{k+1}) = \{cm_{k+1}, g_{k+1} \neq 1, n_{k+1}, k2_{k+1}, 0, 0, l_{k+1}, ch_{k+1}\},$$

і такій послідовності можна поставити у відповідність послідовність слів $S_r \to S_{r+1} \to S_{r+2}$ другого речення T_2, таку, що $\widehat{S}_r \equiv \widehat{S}_k$, $\widehat{S}_{r+2} \equiv \widehat{S}_{k+1}$, з характеристиками

$$Z_r(S_r) = \{1, g_r \neq 1, n_r, k2_r, 0, 0, l_r, 2\},$$
$$Z_{r+1}(S_{r+1}) = \{9, g_{r+2}, 0, 0, 0, 0, 0, 0\},$$
$$Z_{r+2}(S_{r+2}) = \{cm_{k+1}, g_{r+2} \neq 1, n_{k+1}, k2_{k+1}, 0, 0, l_{k+1}, ch_{r+2}\}$$

відповідно, то речення T_1 та T_2 тотожні за змістом.

Наприклад, речення «*Цукор важить три кілограми*» і «*Цукор вагою в три кілограми*». Логіко-лінгвістичні моделі речень мають вигляд:

Важить (цукор, три, кілограми), $P_1(x_1, x_2, x_3)$,

Вагою (цукор, три, кілограми), $\widehat{P}_1(x_1, x_3, x_2)$.

Таким чином, предикати та предикатна змінна (аргумент) при такому виді синонімії виступають спільнокореневими словами і тотожні у логіко-лінгвістичних моделях.

4. Ще одним явищем синонімії вважається використання у реченнях природної мови конверсивів, які описують ту саму ситуацію з різних поглядів.

Якщо в першому реченні T_1 знайдено послідовність слів $S_k \to S_{k+1} \to S_{k+2} \to S_{k+3}$, що володіють характеристиками

$$Z_k(S_k) = \{cm_k, 1, n_k, k2_k, 0, 0, l_k, 1\},$$
$$Z_{k+1}(S_{k+1}) = \{5, 0, n_k, k2_k, t_{k+1}, h_{k+1}, l_k, 2\},$$
$$Z_{k+2}(S_{k+2}) = \{cm_{k+2}, g_{k+2} \neq 1, n_{k+2}, k2_{k+2}, 0, 0, l_{k+2}, 3\},$$
$$Z_{k+3}(S_{k+3}) = \{cm_{k+3}, g_{k+3} \neq 1, n_{k+3}, k2_{k+3}, 0, 0, l_{k+3}, 3\},$$

і такій послідовності можна поставити у відповідність послідовність слів $S_r \to S_{r+1} \to S_{r+2} \to S_{r+3} \to S_{r+4}$ другого речення T_2, таку, що $S_r \equiv S_{k+3}$, $\overleftrightarrow{S}_{r+1} \equiv \overleftrightarrow{S}_{k+1}$, $S_{r+3} \equiv S_{k+2}, S_{r+4} \equiv S_{k+3}$ (де $\overleftrightarrow{S}_{r+1} \equiv \overleftrightarrow{S}_{k+1}$ – конверсиви) з характеристиками

$$Z_r(S_r) = \{cm_{k+2}, 1, n_{k+2}, k2_{k+2}, 0, 0, l_{k+2}, 1\},$$
$$Z_{r+1}(S_{r+1}) = \{5, 0, n_r, k2_r, t_{k+1}, h_{k+1}, l_r, 2\},$$
$$Z_{r+2}(S_{r+2}) = \{9, g_{r+3}, 0, 0, 0, 0, 0, 0\},$$
$$Z_{r+3}(S_{r+3}) = \{cm_k, g_{r+3} \neq 1, n_k, k2_k, 0, 0, l_k, 3\},$$
$$Z_{r+4}(S_{r+4}) = \{cm_{k+3}, g_{k+3} \neq 1, n_{k+3}, k2_{k+3}, 0, 0, l_{k+3}, 3\}$$

відповідно, то речення T_1 та T_2 тотожні за змістом.

Такий запис показує, що лівий та правий трансформи описують ситуацію того самого класу. В результаті використання таких синонімічних конструкцій, зміщується логічний акцент, змінюються синтаксичні позиції лексичних морфем, проте зберігається зміст.

Наприклад, речення «*Сергій подарував батькові машину*» і «*Батько отримав від Сергія машину*». Логіко-лінгвістичні моделі речень мають вигляд:

Подарував (Сергій, батькові, машину), $P_1(x_1, x_2, x_3)$,

Отримав (батько, Сергія, машину), $\overleftrightarrow{P}_1(x_2, x_1, x_3)$.

У логіко-лінгвістичних моделях таких синонімічних конструкцій предикатна змінна (суб'єкт) виступає предикатною змінною (об'єктом) у другому реченні і навпаки, а предикат формується за схемою: $P_1 = \overleftrightarrow{P}_1$ – конверсиви.

Висновки і пропозиції. На основі запропонованих формальних умов виявлення синонімічних конструкцій у реченнях природної мови вдалося провести статистичну оцінку способів взаємозаміни у різних типах текстових документів (рис. 2).

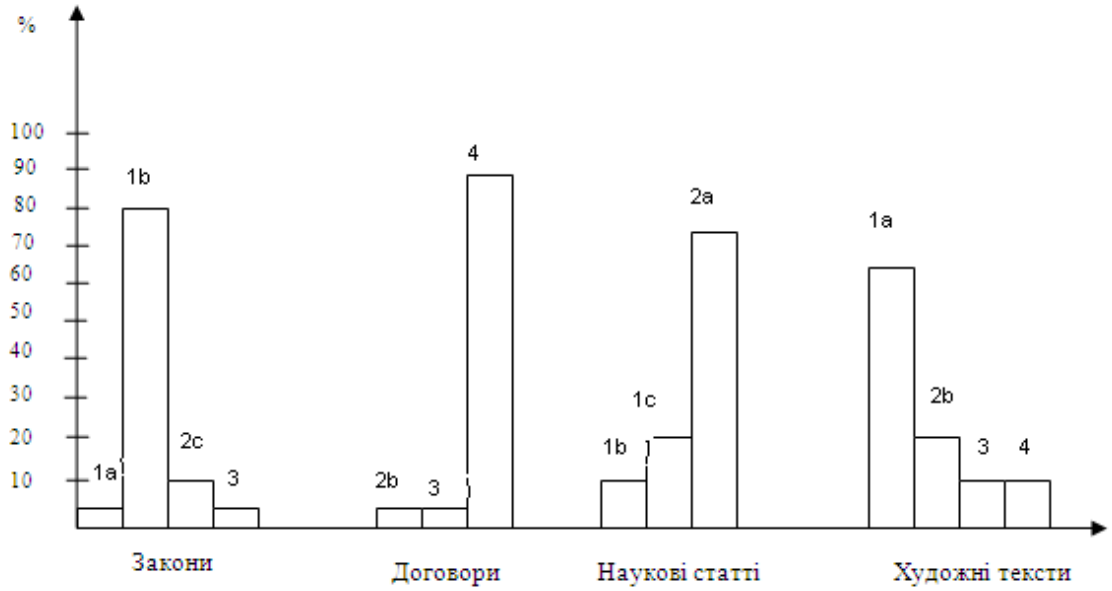

Рис. 2. Статистика використання способів взаємозаміни для різних типів електронних документів

Трансформації, що відбуваються з реченнями природної мови, прямо пропорційно впливають на тип тексту, у якому посилання для кожної умови вживається.

Таким чином, використання запропонованих формальних умов виявлення взаємозамінних синтаксичних конструкцій дає змогу в подальшому оптимізувати процес пошуку синонімів у різних типах документів, а також аналізувати текстову інформацію за змістом.

Список використаних джерел

1. *Апресян Ю. Д.* Исследования по семантике и лексикография. Т. 1 : Парадигматика / Ю. Д. Апресян. – М. : Языки славянских культур, 2009. – 568 с.
2. *Алефиренко Н. Ф.* Спорные проблемы семантики : [монография] / Н. Ф. Алефиренко. – М. : Гнозис, 2005. – 326 с.
3. *Geeraerts Dirk.* Cognitive linguistics: basic readings research / Dirk Geeraerts, Rene Dirven, John R. Taylor. – Berlin–New York : Mouton de cruyter, 2006. – 486 p.
4. *Никитин М. В.* Курс лингвистической семантики : учебное пособие / М. В. Никитин. – 2-е изд. – СПб. : Изд-во РГПУ им. А. И. Герцена, 2007. – 819 с.
5. *Кронгауз М. А.* Семантика / М. А. Кронгауз. – М. : Академия, 2005. – 352 с.
6. *Апресян Ю. Д.* Лексическая семантика : в 2 т. Т. 1 / Ю. Д. Апресян. – М. : Восточная литература, 1995. – 422 с.
7. Режим доступу : https://www.coursera.org/course/nlp.
8. *Алгоритмы:* построение и анализ / Т. Кормен, Ч. Лейзерсон, Р. Ривест, К. Штайн. – 3-е изд. – СПб. : Вильямс, 2013. – 1328 с.
9. *Вавіленкова А. І.* Теоретичні основи аналізу електронних текстів : [монографія] / А. І. Вавіленкова, Д. В. Ланде, О. Є. Литвиненко. – К. : НАУ, 2014. – 250 с.
10. *Гасфилд Д.* Строки, деревья и последовательности в алгоритмах / Д. Гасфилд. – СПб. : Невский диалект БВХ-Петербург, 2003.
11. *Вавиленкова А. И.* Извлечение смысла из предложений естественного языка / А. И. Вавиленкова // Программные продукты и системы. – Тверь : Главная редакция международного журнала НИИ “Центрпрограммсистем”. – 2012. – № 4(100). – С. 87–90.