УДК 004.89
DOI: 10.15587/1729-4061.2018.149596

Аналіз розробленого квантитативного методу автоматичного визначення автора україномовного текстового контенту науково-технічного спрямування

В. В. Литвин, В. А. Висоцька, П. Я. Пукач, З. М. Нитребич, І. І. Демків, А. П. Сеник, О. М. Маланчук, С. І. Саченко, Р.А. Ковальчук, Н. М. Гузик

Запропоновано формальний підхід реалізації визначення автора україномовного тексту. Дослідження проводилось в україномовних наукових текстах технічного профілю. Проаналізовані результати застосування розроблених алгоритмів автоматичного визначення автора текстового контенту на основі методів NLP та стилеметрії. Розглянуто перспективи та особливості застосування інформаційних технологій стилеметрії для визначення автора текстового контенту. Квантитативний контент-аналіз текстового контенту науково-технічного спрямування використовує переваги контент-моніторингу та контент-аналізу тексту на основі методів NLP, Web-Mining та стилеметрії для визначення множини авторів, стилі мовлення яких подібні з досліджуваним уривком тексту. Це звужує коло пошуку при подальшому використанні в методах стилеметрії для визначення ступеня приналежності аналізованого тексту конкретному авторові.

Проведено декомпозицію методу визначення автора на основі аналізу таких коефіцієнтів мовлення як лексична різноманітність, ступінь (міра) синтаксичної складності, зв'язність мовлення, індекси винятковості та концентрації тексту. Паралельно проаналізовані такі параметри авторського стилю як кількість слів у певному тексті, загальна кількість слів цього тексту, кількість речень, кількість прийменників, кількість сполучників, кількість слів із частотою 1, та кількість слів із частотою 10 та більше. Подальшого експериментального дослідження потребує апробація запропонованого методу для визначення ключових слів з інших категорій текстів – наукових гуманітарного профілю, художніх, публіцистичних тощо

Ключові слова: NLP, контент-моніторинг, стоп-слова, контент-аналіз, статистичний лінгвістичний аналіз, квантитативна лінгвістика

1. Вступ

Схема поєднання методів визначення автора україномовного текстового контенту науково-технічного спрямування показує, яка складається з лексичних і синтаксичних рівнів [1]. Використання синтаксичного рівня передбачає обчислення лінгвістичних співвідношень у сполученнях слів [2]. У роботі [3] запропоновано модель для побудови профілю авторського стилю, яка складається з характерного авторського словника та авторського синтаксису [4]. Для опису синтаксису необхідно використати формалізований опис лінгвістичних зв'язків між лексичними одиницями фрази на теоретико-

множинній мові [5]. У роботі [6] висувається формалізований опис будь-якого тексту, але формалізований опис лінгвістичних зв'язків між лексичними одиницями не оновлюється. Формалізований опис тексту також міститься в посиланнях [7]. У довіднику [8] була складена формалізована текстова презентація для автоматизації процедур аналізу наукових та освітніх текстів з метою визначення семантично значущих фрагментів [9]. У роботі [10] викладено теоретико-множинний опис лінгвістичних відносин у фразах. Такі моделі можуть бути використані для опису зображень авторського словника та авторської синтаксії, але вони не враховують статистичну інформацію про частоту словника та синтаксію [11]. Формалізоване опис, яке використовувалось для аналізу тексту термінологічного словника з метою побудови семантичної мережі його термінів, викладено у довіднику [12]. Проте запропонована модель також не передбачає обліку статистичної інформації про частоту виникнення словника та синтаксису [13].

У роботах [1–5] запропоновано та досліджено методи визначення автора україномовного текстового контенту науково-технічного спрямування. Для реалізації цих методів можна використати різні алгоритми [14], зокрема квантитативні [15]. Тому виникає задача аналізу таких алгоритмів з метою пошуку найефективнішого [16].

Авторифікація авторства – це техніка визначення автора тексту, коли неоднозначно, хто її написав [17]. Це корисно, коли декілька людей претендують на авторство однієї публікації [18] або у випадках, коли ніхто не претендує на авторство текстового контенту [19], наприклад, так звані тролі в соціальних мережах під час інформаційної війни [20]. Складність проблеми авторського тексту, очевидно, експоненціально вища, більша кількість вірогідних авторів [21]. Наявність авторських текстових зразків також є суттєвою при просуненні цієї проблеми [22]. Атрибуція авторського тексту включає наступні три проблеми [23]:

– виявлення автора текстового автора з групи імовірних або очікуваних авторів, де автор завжди знаходиться у групі підозрюваних [24];

– не ідентифікація автора текстового автора з групи вірогідних або очікуваних авторів, де автор може не бути в групі підозрюваних [25];

– оцінка можливості даного тексту, написаного даним автором чи ні [26].

Тому задача автоматичного визначення автора текстового контенту науково-технічного спрямування є актуальною й потребує нових (досконаліших) підходів до її розв'язування [27].

2. Аналіз літературних даних і постановка проблеми

Атрибуція тексту – дослідження тексту з метою встановлення авторства або отримання будь-яких відомостей про автора і умови створення текстового документа [17]. Задачі атрибуції поділяють на ідентифікаційні і діагностичні [18]. Ідентифікаційні задачі дають можливість здійснити перевірку авторства [19]:

– підтвердити/виключити авторство певної особи [20];

– перевірити той факт, що автором всього тексту є одна людина [21];

– перевірити, чи автор тексту є при цьому його справжнім автором [22].

Ідентифікаційні задачі вирішують з припущенням, що автор тексту відомий [23]. Діагностичні завдання дозволяють визначити особистісні характеристики автора (освітній рівень, рідна мова, походження, знання іноземних мов, місце постійного проживання тощо) і/або факт свідомого спотворення письмової мови [24]. Діагностичні задачі вирішують з припущенням, що автор тексту невідомий [25]. У цих випадках зазвичай неможливо порівняти досліджуваний текст з текстами автора [26]. Методи атрибуції дозволяють досліджувати текст на п'яти рівнях [27]:

– Пунктуаційний (особливості вживання автором знаків пунктуації, характерні помилки) [28].

– Орфографічний (характерні помилки в написанні слів) [29].

– Синтаксичний (особливості побудови речень, перевагу тих чи інших мовних конструкцій, вживання часів, активного або пасивного застави, порядок слів, характерні синтаксичні помилки) [30].

– Лексико-фразеологічний (словниковий запас автора [31], особливості використання слів і виразів [32], схильність до вживання рідкісних і іноземних слів, діалектизмів, архаїзмів, неологізмів, професіоналізмів, арготизмів [33], навички вживання фразеологізмів, прислів'їв, приказок, «крилатих виразів» тощо) [34].

– Стилістичний (жанр [35], загальну структуру тексту [36], для літературних творів – сюжет [37], характерні зображальні засоби (метафора, іронія, алегорія, гіпербола, порівняння) [38], стилістичні фігури (градація, антитеза, риторичне питання тощо) [39], інші характерні мовні прийоми) [40].

Під «авторським стилем» зазвичай розуміються останні три рівня. Аналіз становить найбільший інтерес і найбільшу складність [41].

Існує досить багато методів аналізу стилю [42]. В цілому є дві великі групи – експертні та формальні [43]. Експертні методи передбачають дослідження тексту професійним лінгвістом-експертом [44]. До формальних відносяться прийоми з теорії ймовірностей і математичної статистики, алгоритми кластерного аналізу та нейронних мереж [46]. Найбільш повна класифікація основних формальних методів атрибуції текстів дана, наприклад, в роботах [1–8, 47]. Як видно, формальні методи найчастіше засновані на порівнянні обчислюваних характеристик текстів, як в теорії розпізнавання образів [49]. Застосування теорії розпізнавання образів у задачі атрибуції текстів можна зустріти, наприклад, в [50]. У загальному випадку текст відображається в вектор обчислених для нього параметрів, кожен з яких об'єктивно характеризує певний набір особливостей тексту [51]. Таким чином, текст графічно відображається в деяку точку n-мірного простору [52]. При такій формалізації автора подають у вигляді аналогічного вектора параметрів – цим вектором є вектор текстів, написаних даними автором [53].

В якості критерію близькості двох текстів обчислюють відстань між відповідними векторами [54]. Набори параметрів та коефіцієнти мовлення подають як звичайні вектори в n-вимірному декартовом просторі з початку координат. Тоді відстанню між текстами є звичайна декартова відстань між кінцями відповідних векторів. Така нормова відстань є інтегральною

характеристикою відмінності текстів. І тексти з великою відстанню з високою ймовірністю належать різним авторам. Таким чином, щоб співставити авторство двох текстів, досить обчислити для них параметри і визначити відстань [55]. Щоб зіставити текст з автором, порівнюються вектори параметрів автора і даного тексту, тобто фактично знову порівнюють два тексти – текст зі свідомо відомим автором (еталонний текст) і текст, авторство якого потрібно встановити, підтвердити або спростувати (аналізований/досліджуваний текст) [56]. Складають також вектори формальних параметрів, що розрізняють не конкретних авторів (або групи), а виділяють певні характеристики авторів (наприклад, освітній рівень) [57]. У більшості випадків в якості характеристичних параметрів тексту обирають статистичні характеристики:

– кількість використання певних частин мови, деяких конкретних слів, знаків пунктуації, фразеологізмів, архаїзмів, рідкісних та іноземних слів,

– кількість і довжина речень (виміряна в словах, складах, знаках), середня довжина речення,

– кількість повнозначних і службових слів,

– обсяг словника, відношення кількості дієслів до загальної кількості слововживань в тексті тощо [58].

Основна проблема формальних методів аналізу авторства полягає якраз у виборі параметрів та коефіцієнтів мовлення [59]. Існує цілий ряд формальних статистичних характеристик текстів, непридатних для визначення авторства в силу одного з двох недоліків [1–5, 60].

– Відсутність стійкості. Розкид значень параметра для текстів одного і того ж автора настільки великий, що діапазони можливих значень для різних авторів перетинаються. Очевидно, даний параметр не допоможе розрізняти авторів, а при використанні в складі групи параметрів лише зіграє роль додаткового інформаційного шуму [61].

– Відсутність здібності розрізняти. Параметр може приймати близькі значення для всіх або більшості авторів, оскільки його значення визначають властивостями мови, на якому написані тексти, а не індивідуальними особливостями автора тексту. Тому параметри повинні попередньо досліджуватися на стійкість і здатність розрізняти, бажано на текстах великої кількості різних авторів [62].

В роботах [3, 63-65] виділені наступні умови застосовності формального коефіцієнта мовлення стилю автора:

– Масовість (використання тих характеристик тексту, які слабо контролюються автором на свідомому рівні, щоб усунути можливість свідомого спотворення автором характерного для нього стилю або імітації стилю іншого автора) [3, 63].

– Стійкість (збереження постійно значення для одного учасника, але деяке відхилення значень від середнього має бути досить малим) [3, 64].

– Здатність розрізняти (приймає істотно різні значення для різних авторів, тобто перевищують коливання, можливі для одного учасника) [3, 65].

Обрати коефіцієнти та параметри мовлення, які гарантовано розрізняють двох будь-яких авторів, дуже важко [66]. Якими б не були параметри, завжди

існує ймовірність того, що два або більше учасника є за даними параметрами близькими в силу випадкового збігу [67]. Тому на практиці є достатнім, щоб параметр дозволяв впевнено розрізняти між собою різні підмножини авторів, тобто існувала б досить велика кількість підмножин авторів, для яких середні значення параметра значно відрізняються [68]. Параметр, очевидно, не допоможе розрізнити тексти авторів з однієї підмножини, але дозволить впевнено розрізняти тексти авторів, які потрапили в різні підмножини [69]. Розрізняти тексти авторів однієї підмножини можна за рахунок використання одночасно досить великого вектора різних за характером параметрів – в цьому випадку ймовірність випадкового збігу стане помітно меншою. Для впевненого виведення текстів, для яких формально обчислена параметрична відстань мала, необхідно провести додаткове дослідження експертними методами, наприклад, аналіз ключових і/або стопових (службових) слів [70].

Отже, виникає необхідність із-за відсутності практичних експериментів визначення стилю автора для україномовних науково-технічних текстів провести дослідження в цьому напряму. Для розв'язку задачі плагіату як копірайту в наш час вже розроблено багато систем. Щодо рерайту – для славянських мов досить складно вирішити таку задачу із наявності великої множини синонімів та можливості перебудови речень з використанням інших закінчень. Це питання не стосується використання службових слів, так як більшість людей при плагіаті навіть не звертає на них увагу. Тому це і спонукає досліджувати задачу ідентифікації стилю автора для визначення ступеню належності конкретного тексту конкретному автору.

3. Мета та завдання дослідження

Метою роботи є аналіз квантитативних алгоритмів автоматичного визначення автора україномовного текстового контенту науково-технічного спрямування на основі технологій стилеметрії та NLP.

Для досягнення мети сформульовані такі завдання:

– розробити метод визначення автора тексту на основі аналізу алгоритмів та коефіцієнтів лексичного авторського мовлення в еталонному тексті;

– розробити програмне забезпечення контент-аналізу для визначення автора в україномовних текстах на основі стилеметричного аналізу коефіцієнтів мовлення текстового контенту;

– здійснити аналіз результатів експериментальної апробації запропонованого методу на основі контент-аналізу для порівняння алгоритмів автоматичного визначення автора в україномовних наукових текстах технічного профілю.

4. Метод визначення стилю автора текстового контенту

В основу розробленого методу покладено декілька алгоритмів.

Алгоритм I. Попереднє опрацювання даних на основі контент-аналізу (парсинг, сегментація та токенізація тексту, а також лінгвістичний аналіз тексту).

Алгоритм II. Розрахунок та аналіз параметрів мовлення кожного

досліджуваного автора (частота вживання слів, кількість знаків пунктуації, символів, речень, слів і співвідношення кількості знаків і речень).

Алгоритм III. Розрахунок та аналіз коефіцієнтів мовлення кожного досліджуваного автора (лексична різноманітність, ступінь синтаксичної складності, зв'язність мовлення, індекси винятковості та концентрації тексту).

Алгоритм IV. Класифікація за цим факторами учасників проекту (використання трьох класифікаторів як нечіткі, SVM і комбінація цих двох).

Алгоритм V. Аналіз продуктивності для визначення точності кожного класифікатора.

Алгоритм VI. Ідентифікація через накладання фільтрів підмножини ймовірних авторів з множини всіх досліджуваних (алгоритми VIII–XI).

Для досягнення мети дослідження розроблено систему типу лексер з можливістю обрання мови/мов аналізованого контенту, яка реалізована на Web-ресурсі Victana [16]. Лексер (токенізатор, сегментатор) – це частина аналізатора тексту на природній мові. Завдання лексера – виділити в тексті основні структурні одиниці – лексеми і розпізнати, зіставивши зі словниковими формами або іншими морфологічними зразками. В результаті роботи лексера виходить складна структура даних – граф токенізації. Граф токенізації є вихідним матеріалом для роботи синтаксичного парсеру (рис. 1). У вузлах графа є маркери. Кожен токен зберігає інформацію про місцезнаходження витягнутого слова в початковому тексті (індекс першого символу і кількість символів в слові), саме слово і результати його ідентифікації. Зліва завжди знаходиться спеціальний токен ідентифікації початку речення. Кожен лист графу є спеціальним токеном закінчення речення. Кожен шлях в графі закінчується спеціальним токеном. Для більшості випадків цей токен позначає праву межу речення. Таким чином, синтаксичний аналізатор (парсер) має можливість враховувати близькість слів до меж висловлювання, що є корисним для оптимізації деяких правил фільтрації токенів.

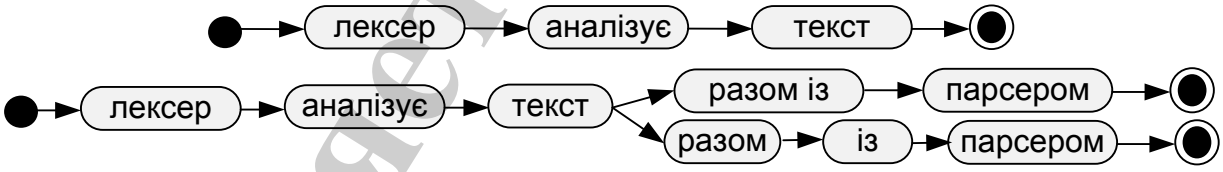


Рис. 1. Приклади графів токенізації для україномовного речення

Лексер працює в дуже тісній взаємодії з парсером тексту. Розпізнані лексером слова підтверджують/спростовують висунуті парсером гіпотези про синтаксичну структуру тексту. Парсер на основі поточного контексту висуває нові гіпотези, які переривають/продовжують окремі шляхи токенізації в графі. Таким чином, токени витягуються під час роботи правил парсера і негайно перевіряються на відповідність умов в цих правилах. Наприклад, без урахування синтаксичних правил ланцюжок «ямаладонькататамого» допускає множину варіантів розбиття на слова (рис. 2).

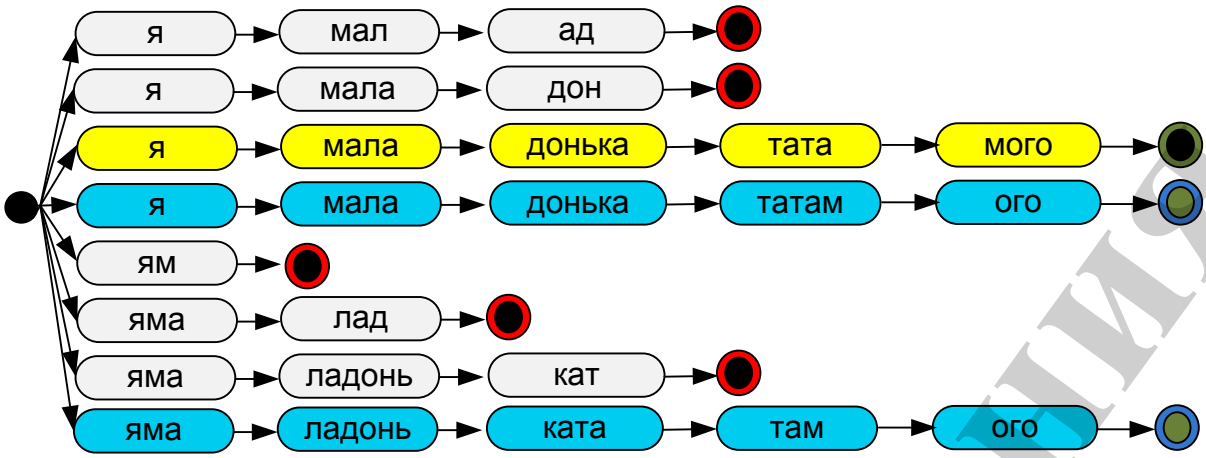


Рис. 2. Приклад графу токенізації без урахування синтаксичних правил

Деякі шляхи перериваються через неможливість знайти в словнику відповідне слово. Простий жадібний алгоритм за правилом «ідентифікувати з вхідного буфера максимально довге слово, знайдене в лексиконі» не працює. Лексер з'являється в системі опрацювання тексту [16] в результаті декомпозиції задачі парсингу. Спрощує реалізацію морфологічного і синтаксичного аналізаторів (рис. 3), так як дозволяє працювати з більшими одиницями – лексемами. Введене таким способом спрощення неявно обмежує спільність всієї системи, так як сама по собі ідея розбиття тексту на незалежні лексеми поєднується не з усіма мовами. Більш того, навіть для мов з природним виділенням слів на листі в звуковому поданні з'являються складні ефекти злиття слів в більші одиниці. У германських мовах це знаходить відображення на листі у вигляді злиття артиклів і прийменників з іншими словами.

Лексер і токенізатор працюють взагалі без явно заданих правил, використовуючи тільки інформацію в лексиконі. Більш-менш обов'язковими є тільки завдання типу кордонів слів у мові і список символів-роздільників.

Додаткові правила допомагають вирішити кілька практичних завдань, збільшуючи ефективність роботи граматичного движка. Зокрема, правила дозволяють зменшувати неоднозначність ідентифікації слів за рахунок часткового зняття омонімії. Алгоритми, які розроблені для вирішення вищеописаних завдань, допускають різне налаштування на об'єктну мову і на особливості опрацьованих текстів та повідомлень. Для кожного виду настройки створені правила. Правила пишуться в текстових вихідних файлах словника відповідно до визначених специфікацій. Специфікації розроблені так, щоб правила можна легко редагувати в будь-якому простому текстовому редакторі або генерувати програмно, наприклад в результаті статистичного опрацювання мовних корпусів. Компілятор словника при трансляції цих специфікацій формально перевіряє коректність правил, оптимізує і зберігає в спеціальному внутрішньому поданні. Потім движок в ході розбору тексту завантажує скомпільовані правила, зазвичай не витрачаючи час на розбір синтаксису (алгоритм VII). Таким чином досягається компроміс між зручністю написання правил і ефективністю використання движком.

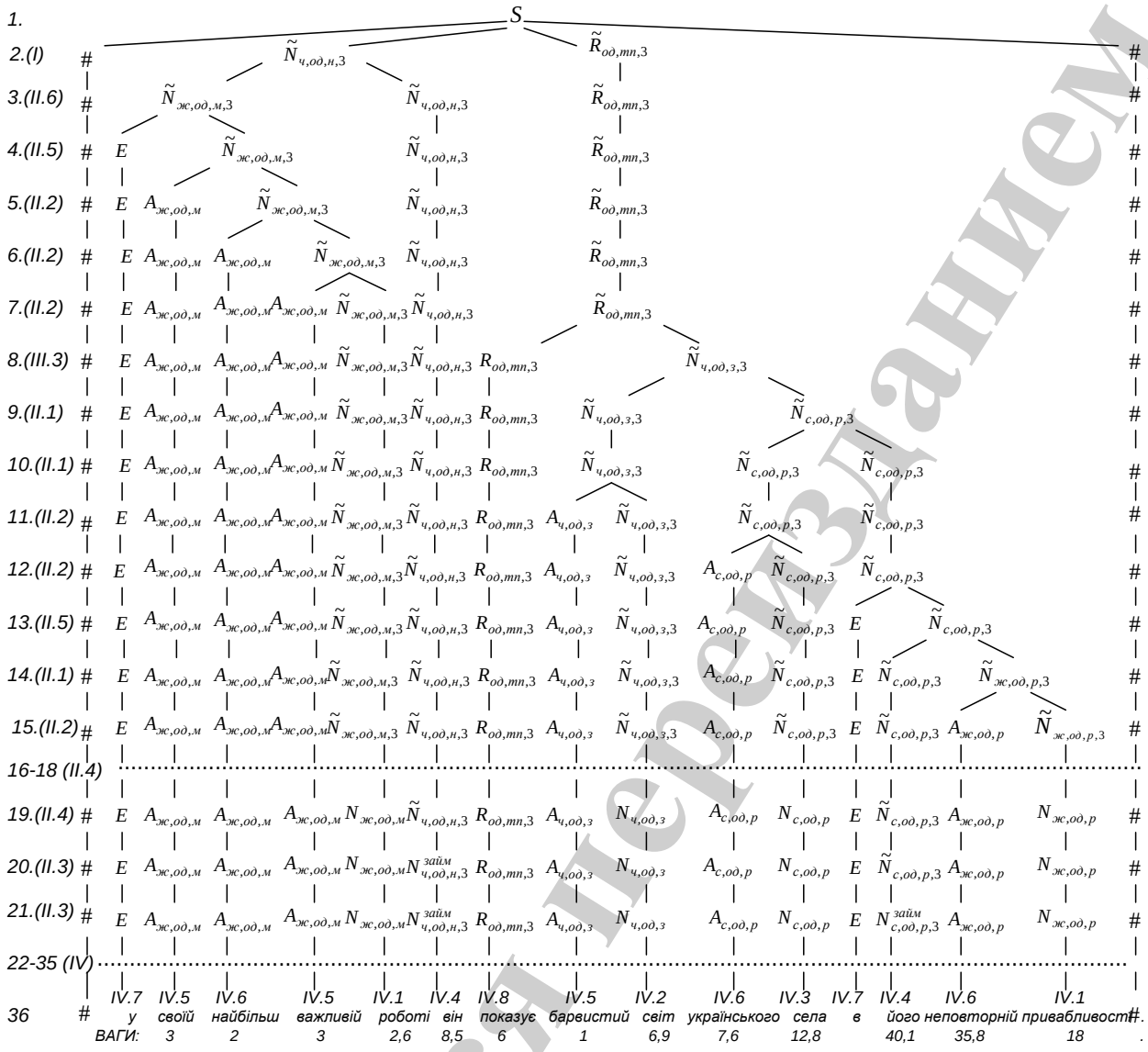


Рис. 3. Результат синтаксичного аналізу до україномовного речення

Алгоритм VII. Сегментатор текстового контенту

Крок 1. Розпізнаванні слова.

Крок 2. Визначення меж лексем.

Крок 3. Визначення повних словоформ.

Крок 4. Ідентифікація неподільних токенів, в яких є точки, пропуски і т.д.

Крок 5. Розбиття тексту на речення.

Символи, які є роздільниками речень (точка, знаки питання та оклику), визначені відповідним параметром в описі мови. Інший параметр в описі мови задає максимальну довжину речення. Використовується для запобігання переповнення внутрішніх буферів і зациклення при розборі складноформатованого тексту, коли алгоритм не може знайти маркер кінця речення. Якщо в якості роздільників використовується точка, то опрацьовується особливим чином, на відміну від знаків ? та !. Справа в тому, що деякі слова можуть містити точку, і це не повинно викликати розриву речення. Типовий приклад – скорочення типу «і т.д.» або абревіатури типу «N.-

Y.». Аналогічно особливо опрацьовуються числа з десяткової крапкою «9.3». Опрацювання таких винятків (токенів з крапкою всередині) спирається на можливість токенізатора розпізнавати в потоці символів спеціальні ланцюжки з роздільниками всередині.

Точка після повної словоформи вважається безумовним роздільником речення. Для цього використано список спеціальних токенів. Якщо після такого токена йде повна словоформа, починається з великої літери і в лексиконі словникова стаття відзначена, що не починається з великої літери, то спеціальний токен є кордоном речення. Наприклад, в тексті «Текст, відео і т.д. Повідомлення, стаття та ін.» перше речення буде відрізано після «т.д.», так як таке слово – Повідомлення – починається з великої літери.

Значення мінімальної довжини повної словоформи використовуються в разі, коли точка йде після повного слова. Так як зазвичай сегментатор дивиться попереду йдуть символи і вважає кордоном речення випадок, коли наступне слово починається з великої літери, то текст «сонце. море. пісок.» буде вважатися одним реченням. Відповідне правило змушує сегментатор перевірити слово перед точкою по лексикону і в разі успіху – вважати точку кордоном речення незалежно від регістру символів наступного слова. Відповідний параметр дозволяє уникнути непотрібних перевірок для випадків «та ін. символи» – задає мінімальну довжину аналізованого повного слова.

Крім визначення меж лексем, лексер також виконує попереднє розпізнавання морфологічних атрибутів слів, перетворюючи лексеми в токени. Для цього лексер використовує інформацію в лексиконі і правила розпізнавання несловникових лексем, а також ряд допоміжних алгоритмів, в тому числі нечітке розпізнавання. При розпізнаванні слова визначаються такі характеристики, як приналежність до певної частини мови і набір граматичних атрибутів. Розрізняють при побудові україномовних речень з прямим порядком слів іменну групу $\tilde{N}$ та дієслівну групу $\check{R}$ (рис. 4, 5) [1–5].

I) $S \to \# \tilde{N}_{РД,ЧЛ,н,ОС}\ \tilde{R}_{ЧЛ,тп,ОС} \#.$

II) $\tilde{N} = \{AN\}$ or $\tilde{N} = N^{p}$

1) $\tilde{N}_{РД,ЧЛ,ВД,З} \to \tilde{N}_{РД,ЧЛ,ВД,З}\ \tilde{N}_{РД',ЧЛ',р,ОС}$;

2) $\tilde{N}_{РД,ЧЛ,ВД,З} \to A_{РД,ЧЛ,ВД}\ \tilde{N}_{РД,ЧЛ,ВД,З}$;

3) $K_1 \tilde{N}_{РД,ЧЛ,ВД,ОС}\ K_2 \to K_1 N^{займ}_{РД,ЧЛ,ВД,ОС} K_2$;

4) $\tilde{N}_{РД,ЧЛ,ВД,З} \to N_{РД,ЧЛ,ВД}$;

5) $\tilde{N}_{РД,ЧЛ,ВД,З} \to E\tilde{N}_{РД,ЧЛ,ВД,З}$;

6) $\tilde{N}_{РД,ЧЛ,ВД,З} \to \tilde{N}_{РД,ЧЛ,ВД,З} \tilde{N}_{РД,ЧЛ,м,З}$.

III) $\tilde{R} = R\tilde{N}$ or $\tilde{R} = \tilde{N}R$

1) $\tilde{R}_{ЧЛ,тп,ОС} \to R_{ЧЛ,тп,ОС}\ \tilde{N}_{РД',ЧЛ',з,ОС'}\ \tilde{N}_{РД'',ЧЛ'',о,ОС''}$;

2) $\tilde{R}_{ЧЛ,тп,ОС} \to R_{ЧЛ,тп,ОС}\ \tilde{N}_{РД',ЧЛ'С,о,ОС'}\ \tilde{N}_{РД'',ЧЛ'',з,ОС''}$;

3) $\tilde{R}_{ЧЛ,тп,ОС} \to R_{ЧЛ,тп,ОС}\ \tilde{N}_{РД',ЧЛ',з,ОС'}$;

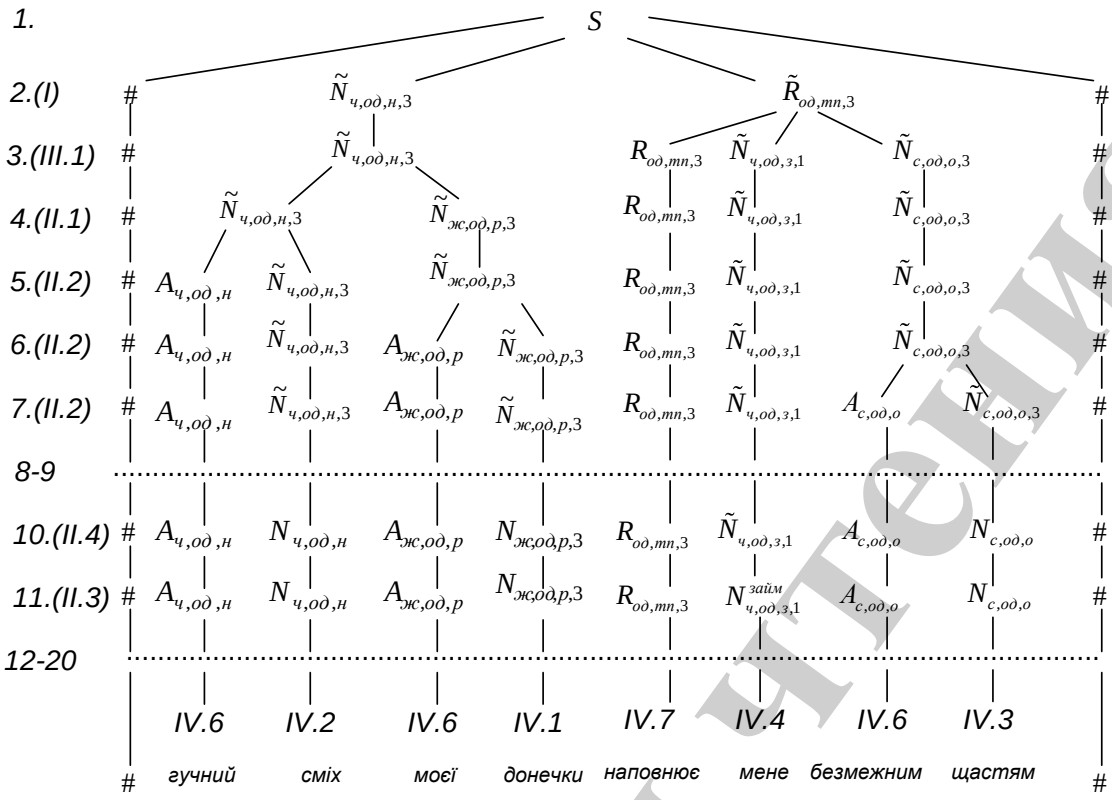
4) $\tilde{R}_{ЧЛ,тп,ОС} \to R_{ЧЛ,тп,ОС}\ \tilde{N}_{РД',ЧЛ',о,ОС'}$;

5) $\tilde{R}_{ЧЛ,тп,ОС} \to R_{ЧЛ,тп,ОС} E\tilde{N}_{РД,ЧЛ,м,З}$:

6) $\tilde{R}_{ЧЛ,тп,ОС} \to E\tilde{N}_{РД,ЧЛ,м,З} R_{ЧЛ,тп,ОС}$.

IV) $Words = \{x_1, x_2, x_3, \ldots, x_n\}$

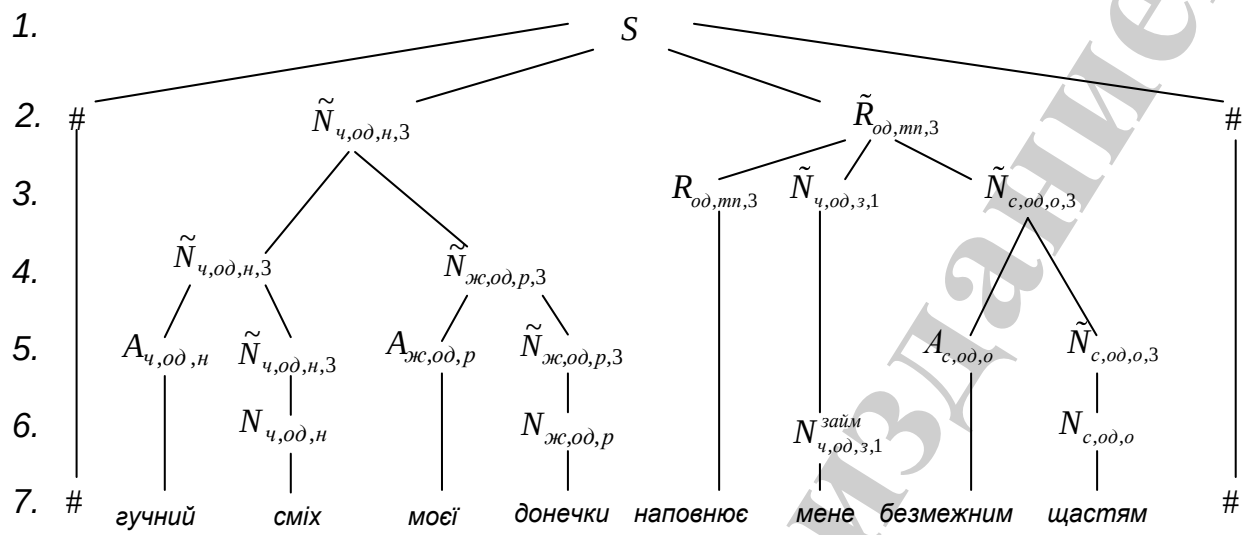
Рис. 4. Правила аналізу україномовного речення, де *А* – прикметник, *N* – іменник, $N^{займ}$ – займенник; число/*ЧЛ* (*од*, *мн*); відмінок/*ВД* (*н*, *р*, *д*, *з*, *о*, *м*, *к*); рід/*РД* (*ч*, *ж*, *с*;), особа/ОС (1, 2, 3); час/*ЧС* (*тп*, *мн*, *мб*)



1. S
2. (I) # $\tilde{N}_{ч,од,н,3}$ $\tilde{R}_{од,тп,3}$ #
3. (III.1) # $\tilde{N}_{ч,од,н,3}$ $R_{од,тп}$ $\tilde{N}_{ч,од,з,1}$ $\tilde{N}_{с,од,о,3}$ #
4. (II.1) # $\tilde{N}_{ч,од,н,3}$ $\tilde{N}_{ж,од,р,3}$ $R_{од,тп}$ $\tilde{N}_{ч,од,з,1}$ $\tilde{N}_{с,од,о,3}$ #
5. (II.2) # $A_{ч,од,н}$ $\tilde{N}_{ч,од,н,3}$ $\tilde{N}_{ж,од,р,3}$ $R_{од,тп}$ $\tilde{N}_{ч,од,з,1}$ $\tilde{N}_{с,од,о,3}$ #
6. (II.2) # $A_{ч,од,н}$ $\tilde{N}_{ч,од,н,3}$ $A_{ж,од,р}$ $\tilde{N}_{ж,од,р,3}$ $R_{од,тп}$ $\tilde{N}_{ч,од,з,1}$ $\tilde{N}_{с,од,о,3}$ #
7. (II.2) # $A_{ч,од,н}$ $\tilde{N}_{ч,од,н,3}$ $A_{ж,од,р}$ $\tilde{N}_{ж,од,р,3}$ $R_{од,тп}$ $\tilde{N}_{ч,од,з,1}$ $A_{с,од,о}$ $\tilde{N}_{с,од,о,3}$ #
8. (II.4) # $A_{ч,од,н}$ $N_{ч,од,н}$ $A_{ж,од,р}$ $\tilde{N}_{ж,од,р,3}$ $R_{од,тп}$ $\tilde{N}_{ч,од,з,1}$ $A_{с,од,о}$ $\tilde{N}_{с,од,о,3}$ #
9. (II.4) # $A_{ч,од,н}$ $N_{ч,од,н}$ $A_{ж,од,р}$ $N_{ж,од,р}$ $R_{од,тп}$ $\tilde{N}_{ч,од,з,1}$ $A_{с,од,о}$ $\tilde{N}_{с,од,о,3}$ #
10. (II.4) # $A_{ч,од,н}$ $N_{ч,од,н}$ $A_{ж,од,р}$ $N_{ж,од,р}$ $R_{од,тп}$ $\tilde{N}_{ч,од,з,1}$ $A_{с,од,о}$ $N_{с,од,о}$ #
11. (II.3) # $A_{ч,од,н}$ $N_{ч,од,н}$ $A_{ж,од,р}$ $N_{ж,од,р}$ $R_{од,тп}$ $N^{займ}_{ч,од,з,1}$ $A_{с,од,о}$ $N_{с,од,о}$ #
12. (IV.6) #*гучний* $N_{ч,од,н}$ $A_{ж,од,р}$ $N_{ж,од,р}$ $R_{од,тп}$ $N^{займ}_{ч,од,з,1}$ $A_{с,од,о}$ $N_{с,од,о}$ #
13. (IV.2) #*гучний сміх* $A_{ж,од,р}$ $N_{ж,од,р}$ $R_{од,тп}$ $N^{займ}_{ч,од,з,1}$ $A_{с,од,о}$ $N_{с,од,о}$ #
14. (IV.6) #*гучний сміх моєї* $N_{ж,од,р}$ $R_{од,тп}$ $N^{займ}_{ч,од,з,1}$ $A_{с,од,о}$ $N_{с,од,о}$ #
15. (IV.1) #*гучний сміх моєї* $N_{ж,од,р}$ $R_{од,тп}$ $N^{займ}_{ч,од,з,1}$ $A_{с,од,о}$ $N_{с,од,о}$ #
16. (IV.7) #*гучний сміх моєї донечки* $R_{од,тп}$ $N^{займ}_{ч,од,з,1}$ $A_{с,од,о}$ $N_{с,од,о}$ #
17. (IV.7) #*гучний сміх моєї донечки наповнює* $N^{займ}_{ч,од,з,1}$ $A_{с,од,о}$ $N_{с,од,о}$ #
18. (IV.4) #*гучний сміх моєї донечки наповнює мене* $A_{с,од,о}$ $N_{с,од,о}$ #
19. (IV.6) #*гучний сміх моєї донечки наповнює мене безмежним* $N_{с,од,о}$ #
20. (IV.3) #*гучний сміх моєї донечки наповнює мене безмежним щастям*#

Рис. 5. Приклад аналізу україномовного речення

Користувач лише може спостерігати як отримаємо дерево складових, або синтаксичну структуру аналізованого речення (рис. 6).



S
(1) $N_{ж,од,н}$
(2) *гучний* $N_{ж,од,н}$
(4) *гучний сміх* $R_{од,з}$
(8) *гучний сміх наповнює* $N_{с,од,з}$
(3) *гучний сміх наповнює безмежним* $N_{с,од,з}$
(7) *гучний сміх наповнює безмежним щастям*.

Рис. 6. Приклад синтаксичної структури аналізованого речення

Для словникових лексем також визначається словникова стаття, формою якої є лексема. В алфавітно-частотних словниках через слеш для слова визначені його характеристики (рис. 7), де A – дієслово, великі літери англійського алфавіту – додаткові характеристики дієслова, V – прикметник, маленькі літери англійського алфавіту – характеристики іменника (рис. 6).

Уривок 1	**Уривок 2**	**Уривок 3**
буферизувати/**ABGH**	клавіатурний/**V**	консоль/ij
відформатувати/**AB**	Кобол/**e**	конфігуратор/**efg**
декодувати/**ABGH**	кодек/**efg**	копілефт/**e**
кешувати/**ABGH**	кодер/**efg**	копірайт/**e**
кирилічний/**V**	кодогенератор/**efg**	криптографічний/**V**
кілобайтовий/**V**	кодосумісний/**V**	криптозахищений/**V**
кілобайт/**efg**	комбосписок/**ab**	крос-асемблер/**efg**
кілобітовий/**V**	комутований/**V**	крос-компілятор/**efg**
кілобіт/**efg**	конкатенація/**ab**	кука/**ab**
кілобод/**efg**	консольний/**V**	курсорний/**V**

Рис. 7. Приклад алфавітно-частотного словника

```
Файл  Правка  Вид  Справка
##################################################################
# Групи a b c d o
#
# -- Перша відміна: іменники жіночого та чоловічого та середнього роду
#
# -- Друга відміна: іменники чоловічого роду із закінченням на -ар -ир
#                   наголошені (Мішана група на -ар -ир)
#
# -- Друга відміна: іменники чоловічого роду з чергуванням -і -о
#
# -- Числівники -ять, -сят, -сто
#
#

SFX a Y 235

#
# ОДНИНА (множина перенесена в гр. b)
#
# Спочатку перша відміна
#
# тверда група в Називному відмінку однини з закінченням на -а
# однина
SFX a   а и [^жчшщ]а  # хата  хати  (Р.)
SFX a   а і [^гґкх]а  # хата  хаті  (Д.М.)
SFX a   а у а    # хата  хату  (З.)
SFX a   а ою [^жчшщ]а  # хата  хатою  (О.)
```

Рис. 8. Словник правил морфологічного аналізу іменників

В базі даних збережені правила приведення до основи слова (рис. 9, а), де flag правило ідентифікації типу слова (наприклад, іменникова група, однина), mask – флексії слова (в квадратних дужках – виключення), find – флексії слова в називному відмінку, repl – флексії слова при відмінюванні (рис. 10).

id	ordering	state	flag	type	lang	mask	find	repl
26	26	1	a	SFX	uk	ін	ін	оном
27	27	1	a	SFX	uk	ін	ін	оні
28	28	1	a	SFX	uk	ір	ір	огу
29	29	1	a	SFX	uk	ір	ір	огові
30	30	1	a	SFX	uk	ір	ір	огом
31	31	1	a	SFX	uk	ір	ір	озі
32	32	1	a	SFX	uk	[^л]ід	ід	оду
33	33	1	a	SFX	uk	[^л]ід	ід	одові
34	34	1	a	SFX	uk	[^л]ід	ід	одом
35	35	1	a	SFX	uk	[^л]ід	ід	оді
36	36	1	a	SFX	uk	[^пг]лід	ід	ьоду
37	37	1	a	SFX	uk	[^пг]лід	ід	ьодові
38	38	1	a	SFX	uk	[^пг]лід	ід	ьодом
39	39	1	a	SFX	uk	[^пг]лід	ід	ьоді
40	40	1	a	SFX	uk	[пг]лід	ід	оду
41	41	1	a	SFX	uk	[пг]лід	ід	одові
42	42	1	a	SFX	uk	[пг]лід	ід	одом
43	43	1	a	SFX	uk	[пг]лід	ід	оді
44	44	1	a	SFX	uk	іб	іб	обу
45	45	1	a	SFX	uk	іб	іб	обові
46	46	1	a	SFX	uk	іб	іб	обом
47	47	1	a	SFX	uk	іб	іб	обі
48	48	1	a	SFX	uk	іп	іп	опу

а

id	ordering	state	word	lang
1	1	1	після	uk
2	2	1	між	uk
3	3	1	are	en
4	4	1	and	en
5	5	7	між	uk
6	6	1	been	en
7	7	1	has	en
8	8	1	their	en
9	9	1	any	en
10	10	1	the	en
11	11	1	with	en
12	12	1	таких	uk
13	13	1	їхніми	uk
14	14	1	как	ru
15	15	1	такої	uk
16	16	1	на	uk
17	17	1	на	ru
18	18	1	ними	uk
20	19	1	для	uk
21	20	1	что	ru
22	21	1	или	ru
23	22	1	это	ru
24	23	1	этих	ru

б

Рис. 9. Правила ідентифікації: *а* – основи слова; *б* – службових слів

```
# Іменники із закінченням на –ін з чергуванням -і -о
SFX a ін    ону   ін      # загін загону          (Д.Р.)
SFX a ін    онові ін      # загін загонові  (Д.)
SFX a ін    оном ін       # загін загоном   (О.)
SFX a ін    оні   ін      # загін загоні           (М.)
третій рядок описує
# Іменники із закінченням на –ір з чергуванням -і -о
SFX a ір    огу   ір      # батір батогу           (Д.Р.)
SFX a ір    огові ір      # батір батогові  (Д.М.)
SFX a ір    огом  ір      # батір батогом   (О.)
SFX a ір    озі   ір      # батір батозі           (М.)
дев’ятий рядок описує
# Іменники із закінченням на –ід з чергуванням -і -о
SFX a ід    оду   [^л]ід  # провід проводу (Д.Р.)
SFX a ід    одові [^л]ід  # провід проводові  (Д.)
SFX a ід    одом  [^л]ід  # провід проводом   (О.)
SFX a ід    оді   [^л]ід  # провід проводі  (М.)
```

Рис. 10. Приклад правил ідентифікації основи слова за аналізом флексій

Також в базі даних (рис. 9, *б*) є словник службових слів, тобто слів, які є додатковими параметрами для аналізу особливостей стилю мовлення автора, та врахування при аналізі текстів впливає суттєво на кінцевий результат.

5. Результати досліджень визначення стилю автора в україномовних науково-технічних текстах

Проаналізуємо 4 розроблені алгоритми, щоб визначити оптимальний розроблений нами метод для ідентифікації стилю автора публікації на основі аналізу його колективних робіт.

Алгоритм VIII. Фільтрація множини аналізованих авторських стилів

```
int i=0, j=0;
while (i<4){
    int c1=0, c2=0, cc2=0;
    while (j<94){
        int s=0;
        while (l<12){
            if ((K[i][l]+abs(F[l]-K[i][l]))>A[j][l]) &&
                    ((K[i][l]-abs(F[l]-K[i][l]))< A[j][l])
                s+=1;
            if (l>6) &&
                ((K[i][l]+abs(F[l]-K[i][l]))>A[j][l]) &&
                        ((K[i][l]-abs(F[l]-K[i][l]))< A[j][l])
                cc2+=s;
            l+=1;
            }
        A2[j]=s;
        A3[j]=cc2;
        c1+=s;
        c2+=s;
        j+=j;
        }
    float t1=c1/94, t2=c2/94;
    int filtr1=0, filtr2=0, filtr3=0
    while (j<94){
        if(A2[j]>=t1) filtr1+=1;
        if(A3[j]>=t2) filtr2+=1;
        if (A2[j]>=t1)&&(A3[j]>=t2)filtr3+=1;
        j+=1;
i+=1;
}
```

Масив K[i][l] – параметри та коефіцієнти стилю для 4-ох колективних робіт (рядки 1–4 в табл. 1 – виділено жовтим кольором). Масив A[j][l] – параметри та коефіцієнти стилю для всіх 94-ох авторів – учасників проекту. Масив F[l] – середні значення параметрів та коефіцієнтів стилю для всіх 94-ох авторів. Алгоритм визначає, чи значення параметрів та коефіцієнтів мовлення

стилю *j*-того автора попадає в межі [x_i+$x_{сер}$; x_i–$x_{сер}$] відхилення значень параметрів та коефіцієнтів мовлення стилю *i*-тої колективної роботи. Заповнюються через фільтри два під масиви А2 (автори, значення більшості параметрів та коефіцієнтів подібні на стиль колективу *i*) та А3 (автори, значення більшості лише коефіцієнтів подібні на стиль колективу i). Далі з отриманих попередніх підмасивів накладанням нового фільтру формується нова підмножина авторів (стилі яких більш подібні на колективні – *i*-ту роботу).

Таблиця 1

Результат роботи алгоритму аналізу стилю автора публікації на інформаційному ресурсі Victana [16]

№	*N*	*W*	W_1	W_{10}	*P*	*Z*	*S*	K_l	K_s	K_z	I_{wt}	I_{kt}
1	622	397	305	5	37	42	48	0,64	0,91	0,81	0,77	0,013
2	614	391	287	4	46	69	32	0,64	0,88	0,73	0,73	0,01
3	658	345	241	8	31	59	42	0,52	0,91	1,07	0,7	0,023
4	631,3	377,7	277,7	5,7	38	56,7	40,7	0,6	0,9	0,88	0,73	0,015
5	661,1	402,7	299,7	4,7	44,7	54,7	24,8	0,61	0,89	0,6	0,74	0,012
6	694,5	417,4	313,1	6,4	54,3	58,5	38,1	0,6	0,87	0,62	0,75	0,015
7	691,8	403,4	301,6	7,8	47,8	60	47,8	0,58	0,88	0,79	0,75	0,019
8	682,5	394,2	291	5	49	61	39,7	0,58	0,88	0,74	0,74	0,013
9	733,5	486,5	392	5	50	65	45	0,66	0,9	0,76	0,8	0,01
.....												
29	704,5	412	303,5	5,5	59	47,5	38	0,58	0,86	0,49	0,74	0,013
30	688,8	416,8	321,9	6	49,7	49,3	41,3	0,6	0,88	0,67	0,77	0,016
.....												
94	680	414	314	4	55	62	34	0,6	0,87	0,58	0,76	0,01

В табл. 1 наведені результати аналізу стилю 94 авторів на одноосібних працях (понад 200 одноосібних робіт) технічного спрямування за період 2001–2017 рр. Для кожного автора виведено середньоарифметичне значення кожного коефіцієнта та параметра мовлення на основі аналізу декількох його робіт за цей визначений період. Також проаналізовані стилі 4-х статей одного авторського колективу під № 1–4 (в табл. 1 виділено жовтим кольором), частина авторів яких є в табл. 3 під № 6 та 30 (в табл. 1 виділено синім кольором).

В результаті отримаємо значення, подані в табл. 2 (алгоритм VIII). Стовпці А – це результат аналізу всіх значень векторів коефіцієнтів та параметрів мовлення авторів з табл. 1. Стовпці В – це результат аналізу лише останніх 5 стовпців в табл. 1. Нажаль цей алгоритм надав такі результати, що наведені автори цих робіт малоймовірно самі написали (найкращі результати виділені червоним кольором – і замало, щоб стверджувати, що вони є авторами понад 50 % цих колективних робіт). Хоча з іншого боку цей алгоритм дає гарні результати – зменшуючи на першому етапі визначення авторства кількість авторів (до 34,04 % із загальної кількості учасників проекту). Це необхідно для

подальшої фільтрації через аналіз стопових слів (прийменників та сполучників) та ключових слів, особливості семантики та лексики при побудові речень тощо.

Таблиця 2

Результат роботи алгоритмів I–IV на інформаційному ресурсі Victana [16]

Алго-ритм	Ко-лек-тив	Середнє значення		Автор				Фільтр			%
				6		30		1	2	3	
		A	B	A	B	A	B	A	B		
VIII	1	5.55319	2.3617	3	2	**6**	2	48	39	35	37,2
	2	7.361702	3.21277	6	3	6	3	40	37	25	26,6
	3	7.521277	3.925532	**8**	**5**	5	**5**	58	35	35	37,2
	4	4.148936	1.457447	3	2	3	0	41	43	33	35,1
	$\bar{x}_i$	6,15	2,74	5,0	3,0	5,0	2,5	46,8	38,5	32,0	34,0
IX	1	5.85106	2.75532	5	2	**8**	**3**	53	53	46	48,9
	2	5.6383	2.7234	**6**	**4**	4	**3**	53	56	43	45,7
	3	3.45745	1.04255	3	0	2	0	40	21	15	15,9
	4	6.2766	2.90426	6	**3**	5	2	44	54	41	43,6
	$\bar{x}_i$	5,31	2,36	5,0	2,3	4,8	2,0	47,5	46,0	36,3	38,6
X	1	6.44681	2.6383	**9**	**3**	**6**	**3**	46	55	42	44,7
	2	7.23404	3.39362	**8**	**4**	**8**	3	45	46	34	36,2
	3	6.46809	2.55319	**8**	**4**	**9**	**4**	48	46	39	41,5
	4	7.8516	3.54255	**9**	3	**9**	**5**	53	51	43	45,7
	$\bar{x}_i$	7,00	3,03	8,5	3,5	8,0	3,8	48,0	49,5	39,5	42,0
XI	1	6.31915	2.11702	**3**	2	**8**	**3**	45	35	29	30,9
	2	4.82979	2.14894	**6**	**3**	**6**	2	51	36	30	31,9
	3	5.89362	2.5	**8**	**4**	**9**	**4**	56	42	41	43,6
	4	5.53191	2.58511	**8**	**3**	**7**	2	49	53	43	45,7
	$\bar{x}_i$	5,64	2,34	6,3	3,0	7,5	2,8	50,3	41,5	35,8	38,0

Тоді проаналізуємо другий алгоритм. Суттєво не відрізняється від попереднього, лише умовою в третьому циклі:

```
if ((K[i][l]+V[l])>A[j][l]) && ((K[i][l]- V[l])< A[j][l]) s+=1
```

де V[l] – масив середніх абсолютних значень відхилень точок даних від середнього значення. В результаті отримаємо значення, подані в табл. 2 (алгоритм IX). Отримані результати трохи покращились, але не настільки, щоб стверджувати, що автори під номером 6 та 30 є справжніми авторами колективних робіт 1–4, хоча вони їх точно писали. З іншого боку, трохи збільшилась кількість авторів (до 38,56 % із загальної кількості учасників проекту) з подібністю в стилі мовлення. Тепер проаналізуємо алгоритм X. В алгоритмі 1 також замінимо в третьому циклі умову на таку:

```
if (abs(A[j][l]- K[i][l])>abs(K[i][l]-F[l])) s+=1
```

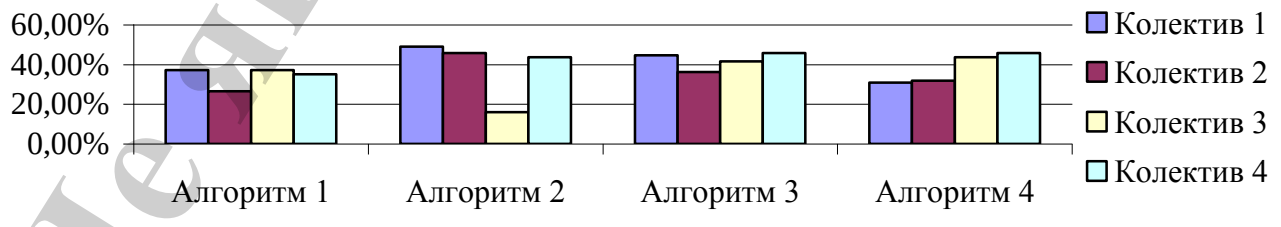
В результаті отримаємо значення, подані в табл. 2 (алгоритм X). Як бачимо – отримані значення гарантовано дають зрозуміти, що стиль авторів під номерами 6 та 30 досить наближений (понад 75–100 %) на стиль колективних робіт 1-4 відповідно (червоним кольором виділені позитивні результати). Хоча значно зросла кількість авторів (до 42,02 % із загальної кількості учасників проекту) з подібністю в стилі мовлення. З іншого боку, в той список багато увійшло тих, то не попав на попередніх етапах дослідження, і випали з множини ті, що увійшли також на попередніх двох етапах дослідження. Тепер спробуємо все таких зменшити ту загальну кількість, застосувавши алгоритм XI до отриманих початкових даних – параметрів та коефіцієнтів мовлення 94-ох учасників проекту. В алгоритмі X вдосконалимо в третьому циклі умову – фільтрацію на таку:

```
if ((abs(A[j][l]- K[i][l])>abs(K[i][l]-F[l])) && (abs(A[j][l]- F[l])>abs(K[i][l]-F[l])))
  || ((abs(A[j][l]- K[i][l])<abs(K[i][l]-F[l])) && (abs(A[j][l]- F[l])<abs(K[i][l]-F[l])))
                        s+=1
```

В результаті отримаємо значення, подані в табл. 2 (алгоритм XI). Отримані значення також підтверджують, що стиль авторів під номерами 6 та 30 досить наближений (понад 75–100 %) на стиль колективних робіт 1–4 відповідно (червоним кольором виділені позитивні результати). Також значно зменшили кількість авторів (до 38,03 % із загальної кількості учасників проекту) з подібністю в стилі мовлення.

6. Обговорення результатів досліджень визначення стилю автора в україномовних науково-технічних текстах

На рис. 11 подані детальні графіки отриманих результатів при застосуванні алгоритмів VIII–XI (під номерами 1–4 відповідно) для аналізу розробеного нами методу визначення стилю автора. На наступному етапі для визначення стилю автора застосовують аналіз стопових слів (прийменників та сполучників) та ключових слів творів авторів, як потрапили до тих 38,03 % (рис. 12). Кожна особистість має свій особливий словниковий запас для передачі своєї думки, в тому числі так званих «паразитичних» (тобто, отже, хоча тощо) та службових слів (і, та, й, але, хоч би тощо).



а

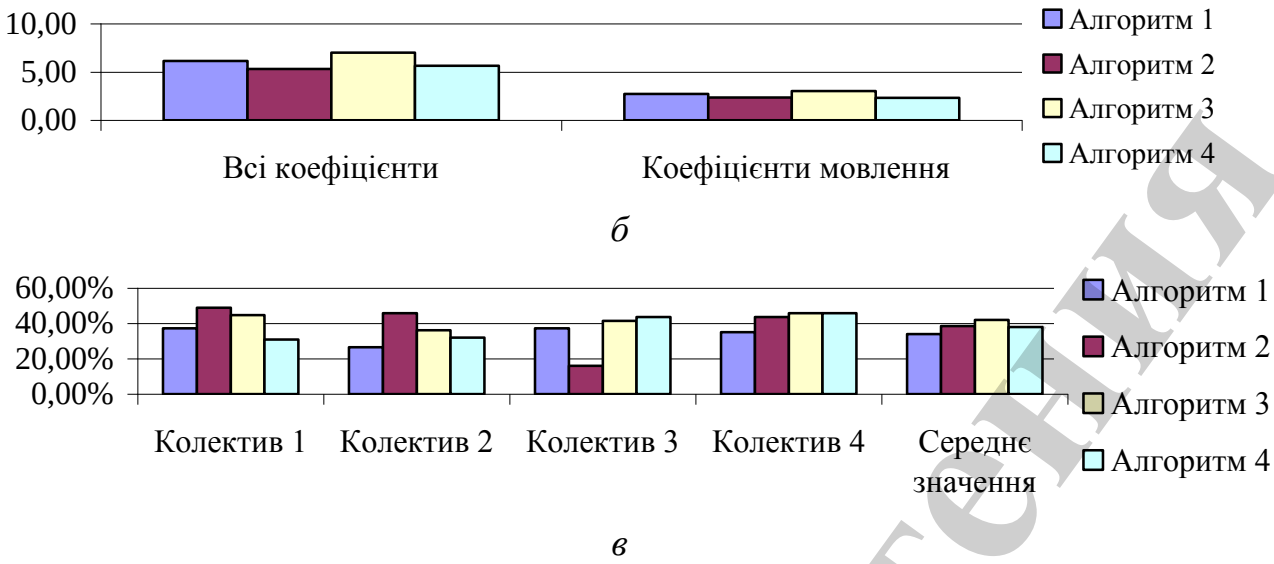


Рис. 11. Детальний аналіз процесу визначення стилю автора:
а – за розробленими алгоритмами; *б* – з врахуванням всіх параметрів та лише коефіцієнтів мовлення; в – для аналізованих колективних робіт

На рис. 12 поданий приклад аналізу стилю автора на другому етапів – через аналіз частоти появи службових та ключових слів з врахування різних фільтрів аналіз повних текстів з списком літератури та анотаціями на різних мова, та аналіз лише інформативної частини публікації, тобто основного тексту з побудовою відповідно частотного словника на 200, 10 та 50 слів).

д *е*

Рис. 12. Детальний аналіз процесу визначення стилю автора на другому етапі для: *а* – повного тексту з побудовою частотного словника зі 100 слів; *б* – основного тексту з побудовою частотного словника зі 100 слів; *в* – повного тексту з побудовою частотного словника з 200 слів; *г* – основного тексту з побудовою частотного словника з 200 слів; *д* – повного тексту з побудовою частотного словника з 50 слів; *е* – основного тексту з побудовою частотного словника з 50 слів

Однак треба зауважити, що є замалою вибірка текстів для аналізу (понад 200) та кількості авторів (94) не гарантує точних результатів. Дослідження має бути продовжене на більшій кількості текстів, до яких, треба зауважити, не завжди маємо доступ. В подальшому необхідно також вдосконалити метод за рахунок аналізу текстів методами стилеметрії та глотохронології.

6. Висновки

1. Розроблено метод визначення автора тексту на основі аналізу коефіцієнтів лексичного авторського мовлення в еталонному уривку авторського тексту. Визначення стилю автора спирається на порівняльний аналіз коефіцієнтів лексичного авторського мовлення: зв'язності мовлення, лексичної різноманітності, синтаксичної складності, індексів концентрації та винятковості для авторського уривку та іншого аналізованого уривку для подальшого порівняння та визначення ступеня належності аналізованого тексту конкретному авторові. Основними стилістичними коефіцієнтами для авторського уривку та іншого аналізованого уривку є зв'язність мовлення, лексична різноманітність, синтаксична складність, а також індекси концентрації та винятковості. Подальший аналіз необхідний для порівняння значень коефіцієнтів та визначення ступеня належності аналізованого тексту конкретному авторові. Особливостями розробленого методу є адаптація морфологічного та синтаксичного аналізу лексичних одиниць до особливостей конструкцій україномовних слів/текстів. Тобто при аналізі лінгвістичних одиниць типу слів, враховувалась належність до частини мови та відмінювання в межах цієї частини мови. Для цього провадився аналіз флексій цих слів для

класифікації, виділення основи для формування відповідних алфавітно-частотних словників. Наповнення цих словників в подальшому враховувалися на наступних кроках визначення авторства тексту як розрахунок параметрів та коефіцієнтів авторського мовлення. Для індивідуального стилю письменника показовими є саме службові (стопові або опорні) слова, оскільки вони ніяк не пов'язані з темою і змістом публікації. Розроблено алгоритм визначення стопових слів текстового контенту на основі лінгвістичного аналізу текстового контенту. Його особливостями є адаптація морфологічного та синтаксичного аналізу лексичних одиниць до особливостей конструкцій україномовних слів/текстів. Наведено теоретичне та експериментальне обґрунтування методу контент-моніторингу та визначення стопових слів україномовного тексту. Метод спрямовано на автоматичне виявлення значущих стопових слів україномовного тексту за рахунок запропонованого формального підходу до реалізації парсингу текстового контенту науково-технічного спрямування.

2. Запропоновано формальний підхід реалізації визначення автора україномовного тексту. Дослідження проводилось в україномовних наукових текстах технічного профілю. Проведено декомпозицію методу визначення автора на основі аналізу таких коефіцієнтів мовлення як лексична різноманітність, ступінь (міра) синтаксичної складності, зв'язність мовлення, індекси винятковості та концентрації тексту. Паралельно проаналізовані такі параметри авторського стилю як кількість слів у певному тексті, загальна кількість слів цього тексту, кількість речень, кількість прийменників, кількість сполучників, кількість слів із частотою 1, та кількість слів із частотою 10 та більше. Проаналізовано розробленою системою понад 200 одноосібних наукових публікацій зі всіх номерів Вісника Національного університету «Львівська політехніка» серії «Інформаційні системи та мережі» за період 2001–2017 рр.

3. Проаналізовані результати застосування розроблених алгоритмів автоматичного визначення автора текстового контенту на основі методів NLP та стилеметрії. Розглянуто перспективи та особливості застосування інформаційних технологій стилеметрії для визначення автора текстового контенту. Квантитативний контент-аналіз текстового контенту науково-технічного спрямування використовує переваги контент-моніторингу та контент-аналізу тексту на основі методів NLP, Web-Mining та стилеметрії для визначення множини авторів, стилі мовлення яких подібні з досліджуваним уривком тексту. Це звужує коло пошуку при подальшому використанні в методах стилеметрії для визначення ступеня приналежності аналізованого тексту конкретному авторові. Проведено порівняння результатів на множині 200 одноосібних робіт технічного спрямування біля 100 різних авторів за період 2001–2017 рр. для визначення чи змінюються і як коефіцієнти різноманітності тексту цих авторів в різні проміжки часу. Отримано експериментальні результати запропонованого підходу для визначення приналежності аналізованого тексту конкретному автору за наявності еталонного інформаційного потоку авторського контенту. Покращують трохи результати відсутність при аналізі вступу та висновків, так як в основному

розкриває свій стиль при описі основної суті свого дослідження. Це досягається за рахунок навчання системи та із перевіркою уточнених заблокованих слів та уточненого тематичного словника.

Література

1. Development of a method for determining the keywords in the slavic language texts based on the technology of web mining / Lytvyn V., Vysotska V., Pukach P., Brodyak O., Ugryn D. // Eastern-European Journal of Enterprise Technologies. 2017. Vol. 2, Issue 2 (86). P. 14–23. doi: https://doi.org/10.15587/1729-4061.2017.98750
2. Analysis of statistical methods for stable combinations determination of keywords identification / Lytvyn V., Vysotska V., Uhryn D., Hrendus M., Naum O. // Eastern-European Journal of Enterprise Technologies. 2018. Vol. 2, Issue 2 (92). P. 23–37. doi: https://doi.org/10.15587/1729-4061.2018.126009
3. Development of a method for the recognition of author's style in the Ukrainian language texts based on linguometry, stylemetry and glottochronology / Lytvyn V., Vysotska V., Pukach P., Bobyk I., Uhryn D. // Eastern-European Journal of Enterprise Technologies. 2017. Vol. 4, Issue 2 (88). P. 10–19. doi: https://doi.org/10.15587/1729-4061.2017.107512
4. The method of formation of the status of personality understanding based on the content analysis / Lytvyn V., Pukach P., Bobyk I., Vysotska V. // Eastern-European Journal of Enterprise Technologies. 2016. Vol. 5, Issue 2 (83). P. 4–12. doi: ttps://doi.org/10.15587/1729-4061.2016.77174
5. Method of functioning of intelligent agents, designed to solve action planning problems based on ontological approach / Lytvyn V., Vysotska V., Pukach P., Vovk M., Ugryn D. // Eastern-European Journal of Enterprise Technologies. 2017. Vol. 3, Issue 2 (87). P. 11–17. doi: https://doi.org/10.15587/1729-4061.2017.103630
6. Khomytska I., Teslyuk V. Specifics of phonostatistical structure of the scientific style in English style system // 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT). 2016. doi: https://doi.org/10.1109/stc-csit.2016.7589887
7. Khomytska I., Teslyuk V. The Method of Statistical Analysis of the Scientific, Colloquial, Belles-Lettres and Newspaper Styles on the Phonological Level // Advances in Intelligent Systems and Computing. Vol. 512. Springer, 2016. P. 149–163. doi: https://doi.org/10.1007/978-3-319-45991-2_10
8. Бук С. Основи статистичної лінгвістики. Львів, 2008. 124 с.
9. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика / Большакова Е., Клышинский Э., Ландэ Д., Носков А., Пескова О., Ягунова Е. Москва: МИЭМ, 2011. 272 с.
10. Анисимов А., Марченко А. Система обработки текстов на естественном языке // Искусственный интеллект. 2002. № 4. С. 157–163.
11. Перебийніс В. Математична лінгвістика. Українська мова. Київ, 2000. С. 287–302.

12. Перебийніс В. Статистичні методи для лінгвістів. Вінниця, 2013. 176 с.
13. Браславский П. И. Интеллектуальные информационные системы. URL: http://www.kansas.ru/ai2006/
14. Ланде Д., Жигало В. Підхід до рішення проблем пошуку двомовного плагіату // Проблеми інформатизації та управління. 2008. № 2 (24). С. 125–129.
15. Варфоломеев А. Психосемантика слова и лингвостатистика текста. Калининград, 2000. 37 с.
16. Victana. URL: http://victana.lviv.ua/nlp/linhvometriia
17. Сушко С., Фомичова Л., Барсуков Є. Частоти повторюваності букв і біграм у відкритих текстах українською мовою // Захист інформації. 2010. Т. 12, № 3 (48). doi: https://doi.org/10.18372/2410-7840.12.1968
18. Когнитивная стилометрия: к постановке проблемы. URL: http://www.manekin.narod.ru/hist/styl.htm
19. Кочерган М. Вступ до мовознавства. Київ, 2005. 368 с.
20. Vysotska V. Linguistic analysis of textual commercial content for information resources processing // 2016 13th International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science (TCSET). 2016. doi: https://doi.org/10.1109/tcset.2016.7452160
21. Родионова Е. Методы атрибуции художественных текстов // Структурная и прикладная лингвистика. 2008. № 7. С. 118–127.
22. Мещеряков Р. В., Васюков Н. С. Модели определения авторства текста // Измерения, автоматизация и моделирование в промышленности и научных исследованиях. 2005. С. 25–29.
23. Морозов Н. А. Лингвистические спектры. URL: http://www.textology.ru/library/book.aspx?bookId=1&textId=3
24. Mobasher B. Data mining for web personalization // The adaptive web. 2007. P. 90–135. doi: https://doi.org/10.1007/978-3-540-72079-9_3
25. Dinucă C. E., Ciobanu D. Web Content Mining // Annals of the University of Petroşani. Economics. 2012. Vol. 12, Issue 1. P. 85–92.
26. Xu G., Zhang Y., Li L. Web content mining // Web Mining and Social Networking. 2011. P. 71–87. doi: https://doi.org/10.1007/978-1-4419-7735-9_4
27. Method of Integration and Content Management of the Information Resources Network / Kanishcheva O., Vysotska V., Chyrun L., Gozhyj A. // Advances in Intelligent Systems and Computing. Vol. 689. Springer, 2017. P. 204–216. doi: https://doi.org/10.1007/978-3-319-70581-1_14
28. Information resources processing using linguistic analysis of textual content / Su J., Vysotska V., Sachenko A., Lytvyn V., Burov Y. // 2017 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS). 2017. doi: https://doi.org/10.1109/idaacs.2017.8095038
29. The risk management modelling in multi project environment / Lytvyn V., Vysotska V., Veres O., Rishnyak I., Rishnyak H. // 2017 12th International

Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). 2017. doi: https://doi.org/10.1109/stc-csit.2017.8098730

30. Peculiarities of content forming and analysis in internet newspaper covering music news / Korobchinsky M., Chyrun L., Chyrun L., Vysotska V. // 2017 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). 2017. doi: https://doi.org/10.1109/stc-csit.2017.8098735

31. Intellectual system design for content formation / Naum O., Chyrun L., Vysotska V., Kanishcheva O. // 2017 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). 2017. doi: https://doi.org/10.1109/stc-csit.2017.8098753

32. The Contextual Search Method Based on Domain Thesaurus / Lytvyn V., Vysotska V., Burov Y., Veres O., Rishnyak I. // Advances in Intelligent Systems and Computing. Vol. 689. Springer, 2017. P. 310–319. doi: https://doi.org/10.1007/978-3-319-70581-1_22

33. Марченко О. Моделювання семантичного контексту при аналізі текстів на природній мові // Вісник Київського університету. 2006. № 3. С. 230–235.

34. Jivani A. G. A Comparative Study of Stemming Algorithms // Int. J. Comp. Tech. Appl. 2011. Vol. 2, Issue 6. P. 1930–1938.

35. Using Structural Topic Modeling to Detect Events and Cluster Twitter Users in the Ukrainian Crisis / Mishler A., Crabb E. S., Paletz S., Hefright B., Golonka E. // Communications in Computer and Information Science. Vol. 528. Springer, 2015. P. 639–644. doi: https://doi.org/10.1007/978-3-319-21380-4_108

36. Родионова Е. Методы атрибуции художественных текстов // Структурная и прикладная лингвистика. 2008. № 7. С. 118–127.

37. Бублейник Л. Особливості художнього мовлення. Луцьк, 2000. 179 с.

38. Kowalska K., Cai D., Wade S. Sentiment Analysis of Polish Texts // International Journal of Computer and Communication Engineering. 2012. Vol. 1, Issue 1. P. 39–42. doi: https://doi.org/10.7763/ijcce.2012.v1.12

39. Kotsyba N. The current state of work on the Polish–Ukrainian Parallel Corpus (PolUKR) // Organization and Development of Digital Lexical Resources. 2009. P. 55–60.

40. Single-frame image super-resolution based on singular square matrix operator / Rashkevych Y., Peleshko D., Vynokurova O., Izonin I., Lotoshynska N. // 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON). 2017. doi: https://doi.org/10.1109/ukrcon.2017.8100390

41. Learning-based image scaling using neural-like structure of geometric transformation paradigm / Tkachenko R., Tkachenko P., Izonin I., Tsymbal Y. // Studies in Computational Intelligence. Vol. 730. Springer, 2018. P. 537–565. doi: https://doi.org/10.1007/978-3-319-63754-9_25

42. Vysotska V., Rishnyak I., Chyryn L. Analysis and Evaluation of Risks in Electronic Commerce // 2007 9th International Conference – The Experience

of Designing and Applications of CAD Systems in Microelectronics. 2007. doi: https://doi.org/10.1109/cadsm.2007.4297570

43. Vysotska V., Chyrun L., Chyrun L. Information technology of processing information resources in electronic content commerce systems // 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT). 2016. doi: https://doi.org/10.1109/stc-csit.2016.7589909

44. Vysotska V., Chyrun L., Chyrun L. The commercial content digest formation and distributional process // 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT). 2016. doi: https://doi.org/10.1109/stc-csit.2016.7589902

45. Content linguistic analysis methods for textual documents classification / Lytvyn V., Vysotska V., Veres O., Rishnyak I., Rishnyak H. // 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT). 2016. doi: https://doi.org/10.1109/stc-csit.2016.7589903

46. Lytvyn V., Vysotska V. Designing architecture of electronic content commerce system // 2015 Xth International Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT). 2015. doi: https://doi.org/10.1109/stc-csit.2015.7325446

47. Vysotska V., Chyrun L. Analysis features of information resources processing // 2015 Xth International Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT). 2015. doi: https://doi.org/10.1109/stc-csit.2015.7325448

48. Application of sentence parsing for determining keywords in Ukrainian texts / Vasyl L., Victoria V., Dmytro D., Roman H., Zoriana R. // 2017 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). 2017. doi: https://doi.org/10.1109/stc-csit.2017.8098797

49. Maksymiv O., Rak T., Peleshko D. Video-based Flame Detection using LBP-based Descriptor: Influences of Classifiers Variety on Detection Efficiency // International Journal of Intelligent Systems and Applications. Vol. 9, Issue 2. P. 42–48. doi: https://doi.org/10.5815/ijisa.2017.02.06

50. Peleshko D., Rak T., Izonin I. Image Superresolution via Divergence Matrix and Automatic Detection of Crossover // International Journal of Intelligent Systems and Applications. 2016. Vol. 8, Issue 12. P. 1–8. doi: https://doi.org/10.5815/ijisa.2016.12.01

51. The results of software complex OPTAN use for modeling and optimization of standard engineering processes of printed circuit boards manufacturing / Bazylyk O., Taradaha P., Nadobko O., Chyrun L., Shestakevych T. // 2012 11th International Conference on "Modern Problems of Radio Engineering, Telecommunications and Computer Science" (TCSET). 2012. P. 107–108.

52. The software complex development for modeling and optimizing of processes of radio-engineering equipment quality providing at the stage of

manufacture / Bondariev A., Kiselychnyk M., Nadobko O., Nedostup L., Chyrun L., Shestakevych T. // TCSET'2012. 2012. P. 159.

53. Riznyk V. Multi-modular Optimum Coding Systems Based on Remarkable Geometric Properties of Space // Advances in Intelligent Systems and Computing. Vol. 512. Springer, 2017. P. 129–148. doi: https://doi.org/10.1007/978-3-319-45991-2_9

54. Development and Implementation of the Technical Accident Prevention Subsystem for the Smart Home System / Teslyuk V., Beregovskyi V., Denysyuk P., Teslyuk T., Lozynskyi A. // International Journal of Intelligent Systems and Applications. 2018. Vol. 10, Issue 1. P. 1–8. doi: https://doi.org/10.5815/ijisa.2018.01.01

55. Basyuk T. The main reasons of attendance falling of internet resource // 2015 Xth International Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT). 2015. doi: https://doi.org/10.1109/stc-csit.2015.7325440

56. Pasichnyk V., Shestakevych T. The model of data analysis of the psychophysiological survey results // Advances in Intelligent Systems and Computing. Vol. 512. Springer, 2017. P. 271–281. doi: https://doi.org/10.1007/978-3-319-45991-2_18

57. Zhezhnych P., Markiv O. Linguistic Comparison Quality Evaluation of Web-Site Content with Tourism Documentation Objects // Advances in Intelligent Systems and Computing. Vol. 689. Springer, 2018. P. 656–667. doi: https://doi.org/10.1007/978-3-319-70581-1_45

58. Chernukha O., Bilushchak Y. Mathematical modeling of random concentration field and its second moments in a semispace with erlangian disrtibution of layered inclusions // Task Quarterly. 2016. Vol. 20, Issue 3. P. 295–334.

59. Davydov M., Lozynska O. Information system for translation into ukrainian sign language on mobile devices // 2017 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). 2017. doi: https://doi.org/10.1109/stc-csit.2017.8098734

60. Davydov M., Lozynska O. Mathematical Method of Translation into Ukrainian Sign Language Based on Ontologies // Advances in Intelligent Systems and Computing. Vol. 689. Springer, 2018. P. 89–100. doi: https://doi.org/10.1007/978-3-319-70581-1_7

61. Davydov M., Lozynska O. Linguistic models of assistive computer technologies for cognition and communication // 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT). 2016. doi: https://doi.org/10.1109/stc-csit.2016.7589898

62. Mykich K., Burov Y. Uncertainty in situational awareness systems // 2016 13th International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science (TCSET). 2016. doi: https://doi.org/10.1109/tcset.2016.7452165

63. Mykich K., Burov Y. Algebraic Framework for Knowledge Processing in Systems with Situational Awareness // Advances in Intelligent Systems and

Computing. Vol. 512. Springer, 2017. P. 217–227. doi: https://doi.org/10.1007/978-3-319-45991-2_14

64. Mykich K., Burov Y. Research of uncertainties in situational awareness systems and methods of their processing // Eastern-European Journal of Enterprise Technologies. 2016. Vol. 1, Issue 4 (79). P. 19–27. doi: https://doi.org/10.15587/1729-4061.2016.60828

65. Mykich K., Burov Y. Algebraic model for knowledge representation in situational awareness systems // 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT). 2016. doi: https://doi.org/10.1109/stc-csit.2016.7589896

66. Kravets P. The control agent with fuzzy logic // Perspective Technologies and Methods in MEMS Design, MEMSTECH'2010 – Proceedings of the 6th International Conference. Lviv, 2010. P. 40–41.

67. On the Asymptotic Methods of the Mathematical Models of Strongly Nonlinear Physical Systems / Pukach P., Il'kiv V., Nytrebych Z., Vovk M., Pukach P. // Advances in Intelligent Systems and Computing. Vol. 689. Springer, 2018. P. 421–433. doi: https://doi.org/10.1007/978-3-319-70581-1_30

68. Kravets P. The Game Method for Orthonormal Systems Construction // 2007 9th International Conference – The Experience of Designing and Applications of CAD Systems in Microelectronics. 2007. doi: https://doi.org/10.1109/cadsm.2007.4297555

69. Kravets P. Game Model of Dragonfly Animat Self-Learning // Perspective Technologies and Methods in MEMS Design (MEMSTECH 2016): Proc. of XII-th Int. Conf. Lviv: Lviv Politechnic Publishing House, 2016. P. 195–201.

70. Vysotska V., Fernandes V. B., Emmerich M. Web content support method in electronic business systems // Proceedings of the 2nd International Conference on Computational Linguistics and Intelligent Systems. Vol. I: Main Conference. Lviv, 2018. P. 20–41. URL: http://ceur-ws.org/Vol-2136/10000020.pdf