## Правова інформатика

УДК 002.6:314.1:004.6

**БРАЙЧЕВСЬКИЙ С. М.,** *кандидат фізико-математичних наук.*

# СТВОРЕННЯ ТЕЗАУРУСІВ НОРМАТИВНО-ПРАВОВОЇ ІНФОРМАЦІЇ В УМОВАХ ЦИФРОВІЗАЦІЇ

***Анотація.*** *В роботі розглядаються питання автоматизації побудови предметних тезаурусів в сфері нормативно-правової інформації. Пропонується використовувати як джерело створення лексикографічного ресурсу документи нормативно-правової бази. Продемонстровані можливості такого підходу на прикладі даних, що стосуються кваліфікації злочинів.*

***Ключеві слова:*** *інформаційні технології, нормативно-правова інформація, тезаурус, кваліфікація злочинів.*

***Summary.*** *The issues of automation of creation of subject thesauri in the field of regulatory and legal information are considered in the work. It is proposed to use documents of the regulatory framework as a source of creating a lexicographic resource. The possibilities of such an approach are demonstrated on the example of data related to the qualification of crimes.*

***Keywords:*** *information technologies, regulatory and legal information, thesaurus, qualification of crimes.*

***Аннотация.*** *В работе рассматриваются вопросы автоматизации построения предметных тезаурусов в сфере нормативно-правовой информации. Предлагается в качестве источника создания лексикографического ресурса использовать документы нормативно-правовой базы. Продемонстрированы возможности такого подхода на примере данных, касающихся квалификации преступлений.*

***Ключевые слова:*** *информационные технологии, нормативно-правовая информация, тезаурус, квалификация преступлений.*

**Постановка проблеми.** Організація та здійснення ефективного парламентського контролю, а також подальшого розвитку національної інтегрованої системи нормативно-правової інформації в умовах цифровізації усіх рівнів інституціального забезпечення державного управління в Україні передбачає активне використання різноманітних автоматизованих систем обробки даних. Одним з важливих напрямків в цій сфері є розробка і створення автоматизованих (а в перспективі автоматичних) засобів своєчасного виявлення найбільш важливих суспільних відносин та встановлення правовідносин [1].

Однією з ключових передумов ефективного функціонування таких засобів є забезпечення структурування розподілених в інформаційних ресурсах даних шляхом формування мереж тематичних текстів. А це, в свою чергу, передбачає формалізоване представлення знань даної предметної області, придатне для автоматизованої обробки. Важливою складовою цієї обробки є автоматизоване створення відповідних онтологій на основі екстрагованих наборів ключових термінів. Створення таких наборів є окремою задачею, що потребує комплексних досліджень. На наш час ця задача на загальному рівні лишається невирішеною [2; 3], а отже, зберігає актуальність.

В пропонованій статті ми обмежимося питанням побудови тезаурусів[*] в семантичному просторі нормативно-правової інформації – базових об'єктів, які широко використовуються для формування баз знань, а також застосовуються в елементах штучного інтелекту.

Звичайно створення таких базових об'єктів містить два окремих етапи: підготовка тематичних масивів текстових документів і власне екстрагування з них ключових термінів та встановлення зв'язків між ними. Як правило, перший етап є доволі складним, оскільки для таких задач не існує універсальних алгоритмів. Визначення тематичної приналежності документа взагалі не є однозначним внаслідок наявності суттєвого суб'єктивного елементу. Тим складніше реалізовувати його на формалізованому рівні. Тому завдання значно спрощується, якщо маємо документальні масиви, які напевно відносяться до потрібної нам тематики.

Нижче покажемо, що одним із шляхів вирішення даної проблеми в сфері нормативно-правової інформації є використання лексичного матеріалу нормативно-правової бази. Він, з одного боку, з достатньою повнотою перекриває правовий термінологічний простір. А з другого боку, є повністю уніфікованим на офіційному рівні, що забезпечує однакове тлумачення в правових текстах різного походження та призначення. Крім того, тексти нормативно-правової бази характеризуються досить низьким рівнем інформаційного "шуму", оскільки містять переважно строгі і точні фахові формулювання. Такий підхід проілюструємо на даних, які стосуються кваліфікації злочинів.

**Результати аналізу наукових публікацій.** Тезауруси широко використовуються в багатьох галузях сучасних інформаційних технологій, пов'язаних з обробкою знань, зокрема і в процесах автоматичної обробки текстів. Розробками в даній сфері займалися вітчизняні й зарубіжні вчені, в тому числі Ланде Д.В., Дмитренко О.О., Радзієвська О.Г. Ягунова Е.Д. Снарський А.А., Проноза Е.В., Лукашевич Н.В., Добров Б.В., Чуйко Д.С. Филиппович Ю.Н., Прохоров А.В., Гладун А.Я., Гендина Н.И., Крижановский А.А., Мозжерина Е.С., Меннінг К.Д. (Manning C.D.), Рагевен П. (Raghavan P.), Шутце Г. (Schütze H.), Селтон Г. (Salton G.), Баклі К. (Buckley C.), Біл Дж. (Beel J.), Гіпп Б. (Gipp B.), Лангер С. (Langer S.), Брейтінгер К. (Breitinger C.). Були досягнуті значні успіхи в контексті побудови тезаурусів "ручними" засобами. Але процес автоматизації створення тезаурусів все ще викликає значні труднощі.

Головна складність полягає в тому, що при цьому доводиться мати справу з семантичними аспектами обробки неструктурованих (недостатньо структурованих) текстів, причому досить специфічних для кожної галузі. Відповідну специфіку доводиться враховувати за допомогою ручної роботи розробників, а також експертних груп. На наш час створені досить розвинені технології побудови предметних тезаурусів. Але вони побудовані на універсальних принципах, що є неспецифічними для тієї чи іншої предметної області.

---

[*] *Від ред.* Тезаурус (від давньогрец. - "скарбниця") – словник найменувань понять та їх зв'язків, призначений для єдиного уніфікованого та формалізованого подання інформації в автоматизованій системі [ДСТУ 2226-93. Автоматизовані системи. Терміни та визначення. [Чинний від 1993-09-09]. Вид. офіц. Київ: Держстандарт України, 1999]. Зазначений словник складається з дескрипторів (стандартизованих ключових слів), в якому їх упорядковано за ієрархією та розміщенням не за алфавітом, а за тематикою груп слів. На відміну від тлумачного словника, тезаурус дозволяє розуміти зміст визначень за допомогою співвіднесення слова з іншими поняттями та їх групами, завдяки чому може використовуватися для наповнення баз знань систем штучного інтелекту.

Тому існує багато окремих задач, майже не розроблених на практичному рівні. Одним з таких напрямків є створення автоматизованих комплексів в галузі нормативно-правової інформації.

**Метою статті** є аналіз та визначення технологічних можливостей використання лексичного матеріалу, що використовується в документах нормативно-правової бази (зокрема на прикладі даних, що стосуються кваліфікації злочинів) для створення тезаурусів, призначених для застосування в автоматизації обробки інформаційних потоків правової інформації.

**Виклад основного матеріалу.** Швидкий розвиток сучасного суспільства породжує нові задачі, серед яких значне місце займає аналіз характеру, спрямованості та повноти інформаційних потоків правової інформації в умовах здійснення децентралізації влади.

Тезауруси є важливою складовою багатьох інформаційних систем, призначених для вирішення різноманітних задач, що передбачають автоматичну обробку текстів [4]. Головне їх значення полягає в тому, що вони орієнтовані на формалізоване представлення знань щодо конкретної предметної області, яка цікавить споживача [5; 6]. Тому є сенс говорити про предметні тезауруси, тобто такі, що пов'язані з певною предметною областю. Основним призначенням предметного тезауруса є представлення стандартизованої термінології, яка використовується для опису інформаційних ресурсів, що відповідають даній предметній області.

Вислів "термінологія" в даному контексті вживається в розширеному розумінні, тобто маються на увазі не лише нормалізовані слова ("терміни" у вузькому розумінні), але й не обов'язково нормалізовані опорні окремі слова (уніграми) та словосполучення, як правило, з двох слів (біграми) і з трьох слів (триграми) [7]. На відміну від словників інших видів, наприклад глосаріїв (які містять лише тлумачення термінів), до складу тезаурусів входять окрім власне термінології відповідні поняття та визначення, що забезпечує можливість формування систем семантичних зв'язків. В деяких випадках тезауруси можуть містити також і окремі семантичні відносини (синоніми, антоніми, пароніми, гіпоніми, гіпероніми тощо).

Побудова тезаурусів, незважаючи на наявність детально розроблених стандартних технологій, є досить складною задачею. Це зумовлює низка причин, з яких для нас головною є відсутність джерел належним чином структурованої лексичної інформації (найбільш поширеним видом такої інформації є розмічені корпуси текстів). Тому для створення тезаурусів спочатку формується відповідний лексикографічний ресурс. Це здійснюється за допомогою тієї або іншої методики, що гуртується на корпусній лінгвістиці [8; 9]. В будь-якому випадку на наш час відсутні автоматичні технології не тільки побудови тезаурусів в повному обсязі, але й формування відповідних лексикографічних ресурсів. Слід також зауважити, що окремою задачею є підтримання тезаурусів в актуальному стані, оскільки лексика, особливо фахова, еволюціонує. Окремою проблемою автоматичних засобів створення тезаурусів є великий обсяг інформаційного "шуму". Причина полягає в тому, що текстові документи, які використовуються, генерувалися з іншою метою і тому, взагалі кажучи, структурно не відповідають задачі побудови тезауруса. Отже, можемо говорити лише про той чи інший рівень автоматизації цих процесів.

Останнім часом популярним стає використання в цій сфері штучного інтелекту [10]. Цей шлях передбачає активне використання роботи експертних груп, що мають на меті керування процесом виведення нових знань з використанням методів самонавчання системи включно з моделюванням міркувань фахівців [11]. Моделювання здійснюється за допомогою набору правил, головними з яких є такі:

- класифікація або розбиття об’єктів на класи, що не перетинаються, за значеннями деякої ознаки;
- порівняння, обчислення різниці у значеннях числових ознак;
- спеціалізація або додавання до опису значень нових ознак;
- діагностика чи виділення значень ознак, що розрізняють задані об’єкти;
- генералізація чи виділення загальної ознаки;
- кластеризація.

У зв’язку з активним розвитком мережних технологій виник новий перспективний напрямок використання предметних тезаурусів як порталів знань [12; 13]. З одного боку, він дозволяє створювати тезауруси як мережні ресурси з відповідними засобами доступу (а отже, і обробки). Тезаурус тоді стає звичайним електронним ресурсом, який може застосовуватись в різноманітних інформаційних системах незалежно від свого створення. З другого боку, він відкриває широкі можливості для використання різноманітних мережних ресурсів як лексикографічної бази при побудові предметних тезаурусів. Однією з головних переваг таких технологій є те, що значна частина мережних інформаційних ресурсів структурується їх генераторами в процесі створення.

Так виникає можливість побудови лексикографічних ресурсів шляхом використання інформаційних масивів, що надаються мережею Інтернет у відкритому доступі. Прикладом може служити побудова корпусів текстових документів в просторі правової інформації шляхом отримання анотацій наукових статей за допомогою пошукової системи Google Scholar (використовувалися запити “Criminal Law” та “Copyright Law”) [1]. Такі мережні ресурси мають достатньо даних практично для повного спектру наявних предметних областей. Особливо слід зазначити високий рівень їх повноти.

Головна проблема у використанні мережних технологій полягає в тому, що використання такого лексикографічного матеріалу вимагає значних обсягів експертної роботи. Значною мірою це пов’язано з тим, що мережні ресурси є різнорідними і містять велику кількість даних, які або взагалі неструктуровані, або частково структуровані, або структуровані відповідно до інших потреб. Тому, як правило, мережні інформаційні ресурси використовуються як сирий матеріал, з якого необхідно створити повноцінну лексикографічну базу.

Отримані текстові документи підлягають подальшій обробці, що включає кілька етапів. Для нас важливим є процес виділення ключових елементарних одиниць тексту (термів), що включає в себе розбиття тексту на токени (елементарні одиниці) та визначення їх вагових множників [14].

Існує ряд методів виділення ключових термів [15]. Найбільш поширеним є класичний статистичний підхід, в якому ваговий множник обчислюється методом TF-IDF [16]. Такий множник є прямо пропорційним частоті вживання терму в даному документі та обернено пропорційним частоті його вживання в усій колекції в цілому. Зауважимо, що частоти нормуються на загальне число слів в документах колекції. Ваговий множник, обчислений за методом TF-IDF, є кількісною мірою важливості слів в окремому документі, що входить до складу колекції [17]. Покажчик TF-IDF також може бути використаний для оцінки релевантності документа по відношенню до певного запиту [18]. Відзначимо, що іноді застосовується спрощений підхід, заснований на використанні лише частоти термів в окремих документах (покажчик TF). Типовим є випадок, коли в специфічній предметній області ключові слова зустрічаються в більшості документів (через тематичну однорідність колекції). Спрощений метод також

використовують для отримання опорних словосполучень (біграм та триграм). Часто в подібних ситуаціях використовують т. з. глобальний покажчик TF-GTF [19].

Для побудови остаточних наборів ключових термів, як правило, здійснюють ще дві операції.

По-перше, з отриманих даних безпосередньо видаляють терми, які з тих чи інших причин не повинні бути присутніми в тезаурусі. Стандартним є метод видалення стоп-слів, що містяться в спеціальних стоп-словниках (окремих для кожної задачі).

По-друге, здійснюється нормалізація термів, що здебільшого зводиться до стематизації, тобто скорочення слова до основи шляхом видалення допоміжних частин, таких як закінчення чи суфікс [20; 21].

Не торкаючись інших, менш важливих для нас технічних питань, скажемо, що на загальному рівні маємо ефективну технологію, що дозволяє здійснювати побудову тезаурусів в відносно простих задачах. Але існують складніші проблеми, з якими доводиться мати справу на практиці.

Застосування в цій сфері сучасних інформаційних технологій супроводжується не лише зростанням обсягів нормативно-правової інформації, але й ускладненням її структури. Це в умовах проведення реформи децентралізації [22] призводить до низки характерних особливостей, які суттєво впливають на ефективність реформ, що проводяться в Україні. Актуальність цих проблем значною мірою викликана тим, що децентралізація відбувається в контексті цифрової трансформації [23] українського суспільства. На цьому шляху виникає низка складних та важливих задач, в тому числі:

- вивчення стану та динаміки функціонування нормативно-правової бази діяльності Верховної Ради України з метою формалізації інформаційних передумов та результатів функціонування комплексної системи парламентського контролю;
- оцінка повноти, періодичності та спрямованості інформаційних потоків забезпечення ефективної діяльності комітету Верховної Ради України з позиції здійснення парламентського контролю;
- формування електронних баз даних нормативно-правових актів в галузі держави і права, як основи створення національної інтегрованої системи нормативно-правових актів в Україні.

Це означає постійне зростання впливу використання цифрових технологій в різноманітних сферах суспільного життя, в тому числі в усьому, що так чи інакше пов’язане з генерацією, обробкою та зберіганням інформації.

Одна з ключових проблем, як ми вже казали, полягає в тому, що значну частину контенту інформаційних потоків нормативно-правової інформації становлять неструктуровані або частково структуровані тексти. Тому виключної ваги набувають технології вичленення окремих фрагментів текстів, що відповідають визначеним семантичним характеристикам.

Проблема структуризації нормативно-правової інформації в широкому розумінні набуває особливої актуальності в сфері використання інформаційних технологій, які передбачають суттєву автоматизацію виробничих процесів. Важлива особливість нормативно-правової інформації полягає в тому, що переважна більшість юридичних термінів лексично не відрізняється (або недостатньо відрізняється) від слів природної мови, які зустрічаються в різноманітних інформаційних ресурсах, що належать до різних предметних областей. Так, наприклад, якщо текст містить кілька вживань слова “кварк”, ми на статистичному рівні можемо вважати його приналежним до наукового твору в галузі фізики. Якщо ж текст містить кілька вживань слова “закон”, ми не маємо жодних підстав вважати його приналежним до правової сфери (“закон природи”, “закон

суспільного розвитку", "закон жанру" тощо). Тому в неструктурованих нормативно-правових текстах визначення опорних слів та словосполучень стає доволі нетривіальною задачею.

Сказане повною мірою стосується і предметних тезаурусів в галузі нормативно-правової інформації, оскільки її особливістю є те, що тексти часто містять лінгвістичні конструкції, які формально (за складом елементів) схожі на фахові, але в дійсності мають інше семантичне наповнення. Це пов'язане з особливістю правової термінології – як фахові терміни в ній вживаються звичайні слова, внаслідок чого неможливо однозначно визначити їх семантичне наповнення. Для нас важливо те, що ця обставина накладає додаткові вимоги на тезауруси, призначені для автоматизації процесів обробки правової інформації.

Додаткова складність полягає в тому, що цифрова трансформація в Україні [23] відбувається в умовах децентралізації влади [24]. Про деякі аспекти цього процесу йшлося в попередній роботі [25]. В ній розглядалася одна з важливих задач цифровізації в сфері нормативно-правової інформації – забезпечення належної її структури. Складність зокрема полягає в тому, що комп'ютерні системи вимагають, щоб інформація, з якою вони працюють, була належним чином формалізована. А цю вимогу важко реалізувати в умовах несинхронізованої діяльності багатьох незалежних агентів інформаційних процесів (термін "агент" в нашому випадку використовується в широкому розумінні, і ми не будемо наводити його строге визначення). Маємо на увазі організації та фізичні особи, які беруть участь в формуванні правової інформації.

В умовах децентралізації одним з ключових чинників є відсутність достатньої синхронізації в процесах генерації правової інформації на різних рівнях адміністративно-територіального устрою України внаслідок суттєвого перерозподілу повноважень між центром і регіонами, а також між різними органами місцевого самоврядування [26 – 28]. Як приклад наведемо дві форми діяльності, що супроводжуються генерацією правової інформації: звітність органів місцевого самоврядування та державний нагляд за законністю рішень органів місцевого самоврядування.

Відсутність єдиних стандартів генерації нормативно-правової інформації в рамках різних видів діяльності на різних рівнях зумовлює неоднорідність її лексики, а також нетотожність семантичного наповнення однотипних лексичних конструкцій. А це значно ускладнює машинну обробку інформаційних потоків.

Сказане вище характеризує особливості процесу автоматизованої побудови предметних тезаурусів в галузі нормативно-правової інформації. Вони стосуються як підготовки лексикографічного ресурсу, так і процесу виділення опорних слів та словосполучень.

Існують різні шляхи вирішення цієї проблеми. На нашу думку, одним із продуктивних підходів є виділення із загальної множини доступних нормативно-правових текстів їх підмножини, яка відповідає таким вимогам:

- тексти гарантовано містять лексику (принаймні основну її частину, достатню для стандартних задач машинної обробки інформаційних потоків), що є типовою для нормативно-правової інформації;
- ця лексика повинна відповідати прийнятим актуальним стандартам юридичної термінології;
- документи повинні бути достатньо структурованими для того, щоб забезпечити можливість виділення потрібних лексичних одиниць в автоматичному режимі;
- контент документів повинен забезпечувати можливість не лише виділення лексичних одиниць, але й визначення зв'язків між ними;
- бажано, щоб документи мали низький рівень інформаційного "шуму".

Перевага нормативно-правової інформації полягає в тому, що така програма, безперечно, може бути реалізована.

Побудова навіть одного предметного тезауруса в повному обсязі є (через згадані вище причини) достатньо складною задачею, і тому наші міркування слід розглядати як основу для майбутнього проекту, здійснення якого потребуватиме значних зусиль і ресурсів.

Нижче ми обговоримо один з можливих підходів, заснований на обробці документальних масивів, що належать до нормативно-правової бази. Перевагою такого підходу є те, що він передбачає використання обмеженого масиву документів, які концентровано містять актуальну правову термінологію і до того ж мають надзвичайно низький рівень інформаційного "шуму". При цьому ми обмежимося ілюстрацією підходу на прикладі даних, що використовуються для кваліфікації злочинів. Звичайно, вони складають обмежений сегмент нормативно-правової інформації, але обсяг лексики, яка міститься в ньому, цілком достатній для відпрацювання базової технології.

Чинне законодавство не містить нормативно-правового визначення поняття "кваліфікація злочинів" [29]. Існують різні способи його розуміння. Для нас найбільш точним є визначення "кваліфікації злочинів" як встановлення тотожності ознак вчиненого суспільно небезпечного діяння і ознак кримінально-правової норми, що передбачає відповідальність за це діяння [30]. Ключову роль тут відіграє те, що вже на базовому рівні фігурують ознаки діяння, тобто те, що передбачає точні офіційні формулювання. Як зазначено в [29], кваліфікація злочинів передбачає встановлення двох важливих обставин:

- факту вчинення суб'єктом злочину суспільно небезпечного діяння, тобто конкретного акту його поведінки (вчинку) у формі дії чи бездіяльності;
- точної відповідності ознак цього діяння ознакам складу злочину, передбаченого відповідною статтею Особливої частини КК України.

Перша обставина однозначно встановлює зв'язок певної дефініції (що фіксується лексичною конструкцією) з конкретним діянням, яке виявляє себе в процесах реального світу. Отже, маємо основу для забезпечення семантичного наповнення опорних слів та словосполучень. Друга обставина забезпечує термінологічну однорідність лексики в усіх текстах нормативно-правової інформації.

Саме ці обставини лежать в основі пропонованого нами підходу до побудови предметних тезаурусів.

Проілюструємо сказане вище прикладами, що ґрунтуються на даних, які містяться в Постанові Пленуму Верховного Суду України "Про судову практику у справах про злочини проти власності" від 06.11.09 р. № 10 [31]. Ми не будемо торкатися конкретних способів виокремлення лексичних конструкцій. Це тема окремого дослідження, що передбачає глибоке вивчення структури документів нормативно-правової інформації. Лише наведемо окремі конкретні слова та словосполучення, які можуть використовуватися як опорні. Візьмемо довільний (далеко не повний) набір лексичних елементів, придатних для побудови опорних слів та словосполучень, пов'язаних з поняттям "крадіжка". Підкреслимо: ми лише ілюструємо основні засади технології, яку ще треба буде створити. Для спрощення викладу будемо також оперувати ненормалізованими елементами.

***Уніграми***

Як уніграми на першому рівні деталізації текстів використовуємо просто терміни, що відповідають вимогам обраного алгоритму (який ми зараз не конкретизуємо), наприклад:

вчинення
підбурювання
готування
крадіжка
грабіж
розбій
викрадення
заволодіння
привласнення
знищення
пошкодження.

***Біграми***

Біграми можуть бути побудовані двома основними шляхами. Вони можуть зустрічатися в явному вигляді, наприклад:

“за умови”
“з метою”
“рухомі речі”
“нерухомі речі”
“грошові кошти”
“цінні метали”
“цінні папери”.

Або вони можуть конструюватися з складніших конструкцій (які являють собою цілісні терміни), наприклад:

“дії майнового характеру” – “майнового характеру”

“електрична та теплова енергія” – “електрична енергія”, “теплова енергія”

“інші діяння щодо певного майна, предметів або засобів” – “інші діяння”, “діяння щодо”, “щодо майна”, “щодо предметів”, “щодо засобів”.

Подібні конструкції можуть також використовуватися для виокремлення уніграм, наприклад:

“інші діяння щодо певного майна, предметів або засобів” – “діяння”, “майно”, “предмети”, “засоби”.

***Триграми***

Побудова наборів триграм аналогічна побудові наборів біграм, наприклад:

“право на майно”
“дії майнового характеру”
“проникнення у житло”

“інші діяння щодо певного майна, предметів або засобів” – “інші діяння щодо”, “діяння щодо майна”, “діяння щодо предметів”, “діяння щодо засобів”.

***Посилання***

Окрему дуже важливу категорію лексичних даних становлять посилання, наприклад:

“ч. 3 ст. 185 КК України”
“ч. 4 ст. 27 – ч. 3 ст. 185 КК України”
“ч. 3 ст. 185; п. 9 ч. 2 ст. 115 КК України”.

Вони не лише можуть виконувати роль опорних словосполучень, але й використовуватися для встановлення зв’язків між різними термінами.

Слід також зазначити, що самі посилання як такі є добре структурованими даними, що суттєво спрощує машинну обробку текстів.

Наведені приклади демонструють можливості створення лексичної бази для побудови предметних тезаурусів, які міститимуть строго визначені терміни, що знаходяться у повній відповідності з правовими нормами. Тим самим значно спрощується робота експертних груп на завершальних стадіях формування власне тезаурусів. Ми також бачимо, що сама структура документів, що стосуються кваліфікації злочинів створює сприятливі умови як для тлумачення окремих термінів, так і для встановлення зв'язків між ними.

**Висновки.**

Таким чином, ми бачимо, що в процесі автоматизації обробки потоків нормативно-правової інформації для створення лексикографічного ресурсу предметних тезаурусів можуть ефективно використовуватися дані нормативно-правової бази. В їх загальній структурі можуть бути виділені (залежно від конкретної поставленої задачі) окремі предметні області, до яких належать документи, що мають добре визначену лексику. Ця лексика містить як фахові юридичні та правові терміни, так і характерні слова та словосполучення, які вживаються в основному в нормативно-правових документах і тому відіграють роль, аналогічну спеціальним термінам.

З іншого боку, документи нормативно-правової бази мають типову (хоча і не завжди формалізовану) структуру, яка дозволяє значно спростити процес аналізу текстів і підвищити ефективність і надійність їх машинної обробки.

На прикладі даних, що стосуються кваліфікації злочинів, була проілюстрована можливість розробки гнучкої методики, яка дозволяє будувати набори опорних слів та словосполучень, придатних для ефективного аналізу текстів і визначення інформаційних та семантичних зв'язків.

Пропоновані нами міркування можуть бути покладені в основу для реального проекту створення автоматизованої системи побудови предметних тезаурусів в семантичному просторі нормативно-правової інформації.

## Використана література

1. Ланде Д.В., Дмитренко О.О., Радзієвська О.Г. Побудова онтологій в галузі права за даними сервісу Google Scholar. *Інформація і право*. № 1(28)/2019. С. 74-85.

2. Лукашевич Н.В., Добров Б.В., Чуйко Д.С. Отбор словосочетаний для словаря системы автоматической обработки текстов: труды международной конференции *Диалог – 2008. Компьютерная лингвистика и интеллектуальные технологии,* г. Москва, 13 мая 2008 г. Москва: РГГУ, 2008. С. 339-344.

3. Филиппович Ю.Н., Прохоров А.В. Семантика информационных технологий: опыты словарно-тезаурусного описания. Москва: МГУП, 2002. 368 с.

4. Лукашевич Н.В. Тезаурусы в задачах информационного поиска. Москва: Изд-во МГУ, 2011. 495 с.

5. Гладун А.Я., Рогушина Ю.В. Онтологии в корпоративных системах. *Корпоративные системы*. Москва: Комиздат, 2006. С. 13-26.

6. Гендина Н.И. Информационно-поисковые тезаурусы: основные виды и области применения. *Научные и технические библиотеки*. Москва: Государственная публичная научно-техническая библиотека России, 2008. С. 5-14.

7. Yagunova E.D. and Lande D.V. Dynamic Frequency Features as the Basis for the Structural Description of Diverse Linguistic Objects. CEUR Workshop Proceedings. Proceedings of the 14th All-Russian Scientific Conference *Digital libraries: Advanced Methods and Technologies, Digital Collections*. Russia, Pereslavl-Zalessky, 2012. Pp. 150-159.

8. Захаров В.П. Корпусная лингвистика: учебно-метод. пособие. СПб, 2005. 48 с.

9. Крижановский А.А. Автоматизированное построение списков семантически близких слов на основе рейтинга текстов в корпусе с гиперссылками и категориями: труды международной конференции *Диалог – 2006. Компьютерная лингвистика и интеллектуальные технологии*, м. Бекасово, 31 мая – 4 июня 2006 г. Москва: РГГУ, 2006. С. 297-302.

10. Мозжерина Е.С. Автоматическое построение онтологии по коллекции текстовых документов: труды 13-й Всероссийской научной конференции *Электронные библиотеки: перспективные методы и технологии, электронные коллекции*. *RCDL'2011,* г. Воронеж 19-22 октяб. 2011 г. Воронеж, 2011. С. 293-298.

11. Naidenova, X.A. Model of Common Sense Reasoning Based on the Lattice Theory. Abstracts of Conference *Mathematical Methods for Learning - 2004. Advances in Data Mining and Knowledge Discovery*, June 21 – 24 2004. Como, Italy. P. 36-39.

12. Загорулько Ю.А., Боровикова О.И. *Автометрия*. Новосибирск, 2008. Т. 44. № 1. С. 100-110.

13. Загорулько Ю.А. Технология разработки порталов научных знаний. *Программные продукты и системы*. 2009. № 4. С. 25-29.

14. Manning C.D., Raghavan P. and Schütze H. An Introduction to Information Retrieval. *Cambridge University Press.* 2009. Pp. 22-36. URL: https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf (дата звернення: 03.12.2021).

15. Lande D.V., Snarskii A.A., Yagunova E.V., and Pronoza E. The Use of Horizontal Visibility Graphs to Identify the Words that Define the Informational Structure of a Text. In: Proceedings of the 12th Mexican International Conference on Artificial Intelligence, 2013. Pp. 209-215. URL: http://dwl.kiev.ua/art/micai/micai-03.pdf (дата звернення: 03.12.2021).

16. Salton G. and Buckley C. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*. № 24(5). 1998. Pp. 513-523. URL: https://www.scinapse.io/papers/1978394996 (дата звернення: 03.12.2021).

17. Ullman J.D. Data Mining, Mining of massive datasets. Cambridge University Press. 2011. Pp. 1-17.

18. Beel J., GIPP B., Langer S., Breitinger C. Research-paper recommender systems: a literature survey. *International Journal on Digital Libraries*. 17(4). 2016. Pp. 305-338.

19. Lande D.V., Dmytrenko O.O., Snarskii A.A. Transformation texts into complex network with applying visability graphs algorithms: материалы XVIII Международной научно-практической конференции *Информационные технологии и безопасность (ИТБ-2018).* Киев: ООО "Инжиниринг". 2018. С. 20-33. CEUR Workshop Proceedings (ceur-ws.org). Vol-2318 urn:nbn: de:0074-2318-4. Selected Papers of the XVIIIШ International Scientific and Practical Conference on *Information Technologies and Security (ITS 2018).*

20. Jongejan B. and Dalianis H. Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike. In the Proceeding of the ACL-2009. Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing. Singapore. August 2-7, 2009. Pp. 145-153.

21. Baeza-Yates R., Ribeiro B. D. A. N. Modern information retrieval. New York: ACM Press. Harlow. England: Addison-Wesle. 2011.

22. Реформа децентралізації. URL: https://www.kmu.gov.ua/diyalnist/reformi/efektivne-vryaduvannya/reforma-decentralizaciyi (дата звернення: 02.12.2021).

23. Цифрова трансформація. URL: https://www.kmu.gov.ua/diyalnist/mizhnarodna-dopomoga/coordination/cifrova-transformaciya (дата звернення: 02.12.2021).

24. Офіційний український державний сайт "Децентралізація влади". URL: https://decentralization.gov.ua (дата звернення: 03.12.2021).

25. Брайчевський С.М. Уніфікація структури правової інформації в умовах децентралізації: матеріали Першої всеукраїнської науково-практичної конференції *Парламентський контроль в умовах децентралізації державної влади та цифрової трансформації в Україні: стан та проблеми*, м. Київ, 30 берез. 2021 р. Одеса: ПП "Фенікс", 2021. С. 40-43.

26. Концепція реформування місцевого самоврядування та територіальної організації влади в Україні. URL: https://zakon.rada.gov.ua/laws/show/333-2014-%D1%80#Text (дата звернення: 23.03.2021).

27. Про внесення змін до Закону України "Про місцеві державні адміністрації" та деяких інших законодавчих актів України щодо реформування територіальної організації виконавчої влади в Україні: проект закону України. URL: http://w1.c1.rada.gov.ua/pls/zweb2/webproc4_1?pf 3511=70293 (дата звернення: 23.03.2021).

28. Нестерович В.Ф. Децентралізація як конституційний принцип здійснення публічної влади на регіональному та місцевому рівнях. *Науковий вісник Дніпропетровського державного університету внутрішніх справ*. 2019. № 3. С. 47-54. URL: https://visnik.dduvs.in.ua/wp-content/ uploads/2019/12/3-19-ua/10.pdf (дата звернення: 23.03.2021).

29. Кваліфікація злочинів; за ред. М.І. Панова. Харків: Право, 2016. 356 с.

30. Навроцький В.О. Основи кримінально-правової кваліфікації. Київ: Юрінком Інтер, 2006. С. 6-44.

31. Про судову практику у справах про злочини проти власності: Постанова Пленуму Верховного Суду України від 06.11.09 р. № 10. URL: https://zakon.rada.gov.ua/laws/show/v00107 00-09#Text (дата звернення: 23.03.2021).

~~~~~~~~~~~~~~ * * * ~~~~~~~~~~~~~~
~~~~~~~~~~~~~~