Міністерство освіти і науки України
Київський національний університет імені Тараса Шевченка
Навчально-науковий інститут філології
Кафедра української мови та прикладної лінгвістики

БАЗА ЗНАНЬ ОСОБОВИХ ІМЕН ЛЮДЕЙ НА ОСНОВІ СЛОВНИКА-ДОВІДНИКА «ВЛАСНІ ІМЕНА ЛЮДЕЙ» Л. СКРИПНИК ТА Н. ДЗЯТКІВСЬКОЇ

Кваліфікаційна робота
на здобуття ОС «бакалавр»
студентки IV курсу,
галузі знань 03 «Гуманітарні науки»,
спеціальності 035 «Філологія»
спеціалізації 035.10 «Прикладна лінгвістика», ОПП «Прикладна (комп'ютерна) лінгвістика та англійська мова»
Анастасії Павлівни МОЗГОВОЇ

Наукові керівники:
д. філософії, доц. **Юлія ЦИГВІНЦЕВА**,
асист. **Валентина РОБЕЙКО**

«Допущено до захисту»
Протокол засідання
кафедри української мови та прикладної лінгвістики
протокол No 16 від «5» 06 2026 року
завідувач кафедри _____(підпис)
канд. філол. н., доц. Сергій РІЗНИК

КИЇВ – 2026

ЗМІСТ

АНОТАЦІЯ

Роботу присвячено створенню бази знань українських особових імен. Актуальність зумовлена тим, що варіантність і непослідовна транслітерація імені ускладнюють його автоматичне опрацювання, а ресурсу, що поєднував би ці відомості, досі бракує.

Об'єктом дослідження є особові імена як складники антропонімічної системи української мови, предметом — структура, варіантність та лексикографічне представлення українських особових імен у форматі бази знань. Мета — створення машиночитної бази знань на основі словника-довідника «Власні імена людей». Робота охоплює три розділи: теоретичний, аналіз джерельної бази та практичну реалізацію бази знань особових імен.

Методологічно поєднано традиційні лінгвістичні методи та сучасні методи комп'ютерної лінгвістики: словник конвертовано у текстовий формат за допомогою оптичного розпізнавання символів, словникові статті опрацьовано засобами регулярних виразів й транслітеровано, а ресурс реалізовано постреляційною базою даних, доцільною для варіативних відомостей про ім'я.

Уперше створено й оприлюднено у відкритому доступі базу знань, що інтегрує варіанти імені, його походження й транслітерацію — відомості, відсутні в наявних джерелах. Ресурс придатний для розв'язання проблем опрацювання природної мови, зокрема «увідповіднення власних імен (named entity matching)», і для наукових досліджень.

Ключові слова: українське особове ім'я, антропоніміка, база знань, транслітерація, опрацювання природної мови, комп'ютерна лінгвістика, словник.

ABSTRACT

The work is devoted to the creation of a knowledge base of Ukrainian personal names. The relevance of the study is determined by the fact that name variation and inconsistent transliteration complicate its automatic processing, while a resource that would consolidate this information has so far been lacking.

The object of the study is personal names as components of the anthroponymic system of the Ukrainian language; the subject is the structure, variation, and lexicographic representation of Ukrainian personal names in knowledge base format. The aim is to create a machine-readable knowledge base founded on the reference dictionary *Vlasni imena liudei*. The work comprises three sections: a theoretical overview, an analysis of the source base, and the practical implementation of the personal names knowledge base.

Methodologically, the work combines traditional linguistic methods with modern computational linguistics techniques: the dictionary was converted into a text format using optical character recognition, dictionary entries were processed by means of regular expressions and transliterated, and the resource was implemented as a post-relational database, well-suited for storing variable name-related data.

For the first time, a knowledge base has been created and published in open access that integrates name variants, etymological origins, and transliteration — information absent from existing resources. The resource is applicable to solving natural language processing challenges, in particular named entity matching, as well as to academic research.

Keywords: Ukrainian personal name, anthroponymy, knowledge base, transliteration, natural language processing, computational linguistics, dictionary.

ВСТУП

Актуальність теми. Особове ім'я виконує насамперед практичну роль — слугує для розрізнення людей. Воно цікавить не лише лінгвістів як мовне явище та джерело наукових студій, а й розробників, оскільки варіантність особових імен спричиняє суттєві труднощі в опрацюванні природної мови.

Попри значний доробок українських ономастів та лексикографів, більшість українських ресурсів описують ім'я лише частково. Зокрема, В. Шульгач створив серію етимологічних антропонімічних словників, П. Чучка розглянув слов'янські особові імена українців з історико-етимологічного погляду. Найвідомішими словниками, що систематизують знання про особове ім’я, є «Словник українських імен» І. Трійняка та словник-довідник «Власні імена людей» Л. Скрипник і Н. Дзятківської, на основі якого і буде формуватися база знань особових імен.

Що ж до електронних ресурсів, у центрі яких було б українське особове ім’я, то аналіз відкритих джерел показав, що в інтернеті можна знайти лише невеликі переліки українських жіночих та чоловічих імен і статистику називання. Відповідно ресурсу, який поєднував би відомості про варіанти імені, його походження та транслітерацію, немає.

Створення бази знань особових імен людей дало б змогу зробити крок до заповнення цієї прогалини: такий ресурс підсилив би рішення для увідповіднення власних імен та інші завдання опрацювання природної мови, а також був би цінним джерелом для дослідників.

Метою роботи є розроблення та верифікація структури бази знань українських особових імен на основі аналізу мікро- та макроструктури словника-довідника «Власні імена людей» Л. Скрипник та Н. Дзятківської, що забезпечить машиночитне представлення відомостей про імена, їхні варіанти, походження та транслітерацію.

Завдання дослідження:

- Розглянути теоретичні засади ономастики й антропоніміки та уточнити термінологічну систему дослідження, зокрема поняття особового імені, його варіантності й транслітерації.
- Висвітлити основні поняття традиційної та комп'ютерної лексикографії, а також теорії баз даних і баз знань, релевантні для опису особових імен.
- Проаналізувати наявні електронні ресурси та відкриті застосунки, що працюють з українськими особовими іменами, визначивши їхні можливості й обмеження.
- Обґрунтувати вибір словника-довідника «Власні імена людей» Л. Скрипник та Н. Дзятківської як джерельної бази дослідження.
- Проаналізувати макро- та мікроструктуру словника-довідника й виокремити розділ, реєстрові одиниці якого слугуватимуть матеріалом для наповнення бази знань.
- Розробити схему бази знань для представлення особових імен, їхніх варіантів, походження та транслітерації.
- Здійснити автоматичне опрацювання джерела (конвертацію, парсинг) та генерацію транслітерацій імен відповідно до розробленої схеми.
- Верифікувати укладену базу знань і забезпечити відкритий доступ до неї.

Об’єктом дослідження є особові імена як складники анпропонімічної системи української мови.

Предметом дослідження є структура, варіантність та лексикографічне представлення українських особових імен у форматі бази знань.

Методологічною основою дослідження слугують положення ономастики й антропоніміки, традиційної та комп'ютерної лексикографії, а також теорії баз даних і баз знань.

Методи дослідження. Для розв'язання теоретичних завдань — опанування засад ономастики, антропоніміки та лексикографії — застосовано аналіз

наукової літератури й описовий метод. Порівняльно-зіставний аналіз використано для огляду наявних джерел і застосунків, обґрунтування вибору словника та порівняння OCR-рішень. Структурно-функціональний аналіз застосовано для дослідження макро- та мікроструктури джерельної бази. Метод комп’ютерного моделювання застосовано для проєктування баз даних.

Наукова новизна роботи полягає у тому, що вперше створено машиночитну базу знань українських особових імен, яка інтегрує в межах єдиного ресурсу офіційні й неофіційні варіанти імені, його походження та транслітерацію — відомості, у наявних джерелах подані лише фрагментарно або відсутні (транслітерацію не містить жодне з них).

Практичне значення. Результати роботи можуть бути використані в прикладних завданнях опрацювання природної мови, насамперед в увідповідненні власних імен людей (named entity matching), а також як джерело для антропонімічних досліджень. Ресурс корисний розробникам україномовного програмного забезпечення, майбутнім батькам і всім, хто цікавиться українськими особовими іменами.

Структура роботи. Кваліфікаційна робота складається зі вступу, трьох розділів із висновками до кожного, загальних висновків, списку використаної літератури (47 позицій) та 1 додатка. Перший розділ закладає теоретичне підґрунтя дослідження: розглянуто особове ім'я в українському онімному просторі та проблему його транслітерації, електронні словники й комп'ютерну лексикографію, а також поняття баз даних і баз знань та огляд наявних ресурсів. Другий розділ присвячено джерельній базі — обґрунтуванню вибору словника «Власні імена людей» та аналізу його макро- й мікроструктури. У третьому розділі описано практичну реалізацію бази знань: конвертацію словника (OCR), парсинг словникових статей, проєктування схеми, формування бази даних особових імен і бази даних транслітерацій та їх об'єднання в єдину базу знань, оприлюднену на GitHub. Загальний обсяг роботи — 58 сторінок, обсяг основного тексту — 50 с.

РОЗДІЛ 1. ТЕОРЕТИЧНІ ЗАСАДИ СТВОРЕННЯ БАЗИ ЗНАНЬ ОСОБОВИХ ІМЕН ЛЮДЕЙ

Перший розділ присвячено теоретичним засадам створення бази знань особових імен. Оскільки така база об'єднує мовний об'єкт, лексикографічне джерело і спосіб організації відомостей, її розробка спирається на теоретичні засади декількох наукових галузей: ономастику й антропоніміку, лексикографію та комп'ютерну лексикографію, а також теорію баз даних і баз знань. Кожному з цих напрямів присвячено окремий підрозділ.

1.1 Особове ім’я в українському онімному просторі

Центральною одиницею цього дослідження є особове ім'я. Воно становить предмет вивчення антропоніміки, яка, своєю чергою, є частиною ономастики. Тому цей теоретичний розділ доцільно розпочати зі стислого огляду обох мовознавчих дисциплін.

1.1.1. Ономастика, класифікація онімів

Відповідно до лексико-граматичних розрядів іменники поділяють на *загальні* та *власні*. Останні відрізняються від загальних тим, що називають предмет як індивідуальний, унікальний [Алексієнко та ін. 2013, с. 251-254]. Власні назви, також відомі як оніми, досліджує мовознавча дисципліна — *ономастика*.

Власні назви генетично можуть виходити з різних частин мови, проте функційно вони виступають як іменники. Існує багато класифікацій онімів і, як зазначає О. Чорноус, класифікація та систематизація онімів повністю не виконана [Чорноус 2021]. Ми послуговуватимемося денотативною та структурною класифікаціями. Енциклопедія «Українська мова» (далі — УМ) пропонує за денотатами розрізняти такі класи онімів: особові імена людей (антропоніми), назви географічних об'єктів (топоніми), назви космічних об'єктів (космоніми), найменування божеств (теоніми), міфічних істот (міфоніми), клички тварин (зооніми), назви організацій, виробничих та суспільних об'єднань

(ергонімія), назви відрізків часу, подій (хрононімія), назви окремих предметів (хрематонімія) [Карпенко 2004, с. 83].

Оскільки об’єктом нашого дослідження є особові імена, то далі сконцентруємо нашу увагу на них. Розділ ономастики, що вивчає особові імена називають *антропонімікою*. Ця дисципліна вирішує не лише теоретичні питання як особливості утворення особових імен, основні принципи номінації людини, виникнення різних форм найменування людини, але і їх правопис, передачу іншою мовою тощо [Желєзняк 2004, с. 29-30].

Антропонімія (сукупність антропонімів) слугує джерелом для багатьох наукових досліджень, наприклад: розселення народів, встановлення фактів асиміляції, соціальне розшарування, вияви класових антагонізмів, розвиток і територіальна концентрація ремесел та торгівлі, організація державно-адміністративних органів тощо [Худаш 1977, с. 13-14].

Українська антропоніміка має тривалу й розгалужену традицію, у межах якої сформувалося кілька самостійних напрямів. Один із них зосереджений на історичному генезисі особових імен — реконструкції найдавнішого шару власних найменувань і з'ясуванні механізмів переходу загальних назв у власні. Так, В. Шульгач здійснив масштабну реконструкцію праслов'янського антропонімного фонду, простеживши становлення особових імен у дописемну добу [Шульгач 2016]. Своєю чергою М. Демчук докладно проаналізувала слов'янські автохтонні особові власні імена, що функціонували в побуті українців XIV-XVII ст., визначивши їхні лексико-семантичні, структурні та етимологічні особливості [Демчук 1988].

Окремий потужний напрям охоплює дослідження прізвищ та історичних процесів становлення сучасної системи іменування особи. М. Худаш вивчав українську антропонімію XVI-XVIII ст., простеживши стабілізацію способів називання людини й обґрунтувавши роль власного особового імені як головного джерела творення українських прізвищ і прізвиськ [Худаш 1977]. Ю. Редько здійснив перше комплексне дослідження, у якому розглянуто виникнення, словотвір та поширення сучасних українських родових назв (прізвищ), а також

уклав їх довідник [Редько 1966]. Регіональний та історико-етимологічний вимір цих студій розширив П. Чучка, який детально опрацював прізвища закарпатських українців і специфіку функціонування родових назв в Українських Карпатах [Чучка 2005].

Значний доробок української антропоніміки пов'язаний також із лексикографічним упорядкуванням особових імен. Праця І. Трійняка «Словник українських імен» [Трійняк 2005] та словник-довідник «Власні імена людей» Л. Скрипник і Н. Дзятківської стали спробою систематизувати інформацію про особове ім'я та його варіанти: у них подаються офіційні імена, уживані в Україні, та їх варіанти, включено відомості про їх походження [Скрипник, Дзятківська 2005]. Етимології антропонімів присвячено багато праць. Зокрема, В. Шульгач докладно описав етимологію антропонімів у серії етимологічних антропонімічних словників [Шульгач 2008-2023], П. Чучка дослідив етимологію слов'янських особових імен українців [Чучка 2011].

На соціолінгвістичному та ареальному ґрунті працювала Л. Кравченко, яка, дослідивши антропонімію Лубенщини, наголосила на здатності неофіційних найменувань відбивати менталітет, експресію та національний колорит мовців [Кравченко 2002].

Термінологічна система антропоніміки ще є не усталеною і серед закордонних та вітчизняних науковців терміни можуть не збігатися та варіюватися [Чорноус 2021]. Як ми зазначили вище, УМ називає антропоніми особовими іменами людей, проте сучасні ономастичні розвідки використовують для усіх груп антропонімів чіткіший термін — власні назви людей, а особовими іменами називають один із їх типів. У науковій роботі ми теж дотримуватимемося цього підходу та називатимемо антропоніми ***власними назвами людей,*** а одиниці *Петро, Марія* тощо — ***особовими іменами*** або ***власними іменами людей***. Усталеної класифікації антропонімів немає, зокрема через те, що антропонімічні моделі варіюються відповідно до культурних особливостей. Однак українські мовознавці до власних назв людей уналежнюють такі групи:

1. особове ім’я, власне ім’я людини *(Олександр, Надія)*;
2. ім’я по батькові *(Григорович, Іванівна)*;
3. прізвище *(Костенко, Синявський)*;
4. прізвисько *(Горішок, Кайдашиха)*;
5. псевдонім *(Леся Українка, Панас Мирний)* [Желєзняк 2004, с. 29]

Оскільки у кваліфікаційній роботі ми будемо працювати лише з особовими іменами, то вважаємо за потрібне детальніше описати цей тип антропонімів.

1.1.2 Особове ім'я серед інших антропонімів

Власне або особове ім’я — це особове найменування, дане дитині при народженні або (рідко) вибране для себе дорослою людиною, наприклад *Надія, Станіслав, Олександр* [Гонца 2021]. Зауважимо, що особове ім’я може бути не лише однослівним, але й багатослівним. Кількість слів в особовому імені залежить від традицій. В Україні, наприклад, поруч з однослівними функціонують подвійні імена такі як: *Анна-Марія, Матвій-Віктор, Ярослав-Петро*. Варто зазначити, що іменам притаманна варіативність, де варіант імені — це розмовно-побутовий (скорочений, чи усічений, здрібніло-пестливий та згрубіло-зневажливий) варіант офіційного повного документального імені [Скрипник, Дзятківська 2005, с. 10]. Останні можуть мати офіційні варіанти, тобто імена, що зустрічаються в документах. Часто ці документальні варіанти імені виходять з розмовно-побутового вжитку і стають офіційними. До прикладу, ім’я *Настя* є офіційним варіантом імені *Анастасія*, проте у розмовно-побутовому вжитку цей варіант широко використовується як скорочений варіант до повного *Анастасія*. Іноді варіанти імені стають офіційними відповідниками через псевдоніми авторів, наприклад особове ім’я *Олесь* стало документальним варіантом імені *Олександр*, оскільки ним часто послуговувалися письменники (наприклад, Олександр Олесь, Олесь Гончар).

1.1.3. Проблема транслітерації українського особового імені

Описана в попередніх підрозділах система українських особових імен функціонує не лише в межах кириличної графіки. В умовах глобалізації

міжнародного документообігу й академічного цитування особове ім'я неминуче потрапляє в латинографічний простір, де постає потреба коректного передавання українського особового імені. Якщо розглянуті вище словники — «Словник українських імен» І. Трійняка та словник-довідник «Власні імена людей» Л. Скрипник і Н. Дзятківської — систематизували офіційні, розмовні та пестливі варіанти імені в межах української мови, то відтворення імені засобами латиниці породжує ще один, формально-графічний, шар його варіантності, що потребує окремого розгляду.

Передавання кириличного письма латинським алфавітом називають романізацією. У її межах розрізняють транслітерацію, що відтворює написання літера за літерою, і транскрипцію, що відтворює звучання слова. Для особових імен, які закріплюються в офіційних документах, визначальною є транслітерація, тоді як транскрипція обслуговує радше потреби фонетичного наближення імені до традицій вимови носіїв інших мов.

Чинною офіційною системою в Україні є Українська національна транслітерація, що ґрунтується на англійській орфографії й потребує лише символів ASCII без діакритиків. Її становлення відбувалося поетапно: першу редакцію закріплено рішенням Української комісії з питань правничої термінології 1996 р., окрему систему для паспортів запроваджено 2007 р., а оновлену версію 2010 р. затверджено Постановою Кабінету Міністрів України № 55 від 27 січня 2010 р. як єдину для всіх власних назв [Постанова КМУ № 55 2010]. Саме національна система визначає сьогодні офіційну, паспортну форму українського особового імені в латинській графіці.

Попри наявність затвердженого стандарту, поза межами закордонного паспорта українські імена та прізвища транслітерують відповідно до нього украй рідко. Так, дослідження англомовного корпусу новинних статей засвідчило, що ім'я того самого українського політика може мати до 21 варіанта транслітерації [Khairova, Mozghova 2025]. Така кількість графічних варіантів імені спричиняє невизначеність у даних і породжує проблему, відому як увідповіднення власних імен людей (англ. *named entity matching*).

Саме тому в нашій роботі ми вважаємо за доцільне доповнити базу знань відомостями про транслітерацію кожного імені та його варіантів. Такої інформації не подає жодне з наявних джерел українських особових імен, проте саме вона здатна закласти підґрунтя для розв'язання завдання увідповіднення власних імен людей, а також інших прикладних завдань опрацювання природної мови. Завдяки цьому створювана база знань стане корисною не лише мовознавцям, а й розробникам.

1.2 Електроні словники та бази знань

Оскільки база знань особових імен людей укладатиметься на основі словника-довідника «Власні імена людей» Л. Скрипник та Н. Дзятківської, доцільно стисло схарактеризувати теоретичні засади лексикографії — дисципліни, що вивчає словники, — а також основні поняття, пов'язані з електронними словниками та комп’ютерною лексикографією.

1.2.1 Основні поняття та методи лексикографії

Перед тим як опанувати термінологічну базу лексикографії, вважаємо за потрібне з’ясувати різницю між лексикографією та словникарством, як такими термінами, що є взаємозалежними, але не еквівалентними.

Термін «лексикографія» трактують по-різному. Зокрема, у навчальному посібнику «Традиційна та комп'ютерна лексикографія» В. Перебийніс та В. Сорокіна: Лексикографія — це галузь мовознавства, предметом якої є методи та практика укладання словників [Перебийніс, Сорокін, с. 4]. Словникарство ж — це власне практика укладання словників. Також у цей термін входять словники певної мови різного типу і статусу, а також словникова індустрія [Демська 2010, с. 22-23].

Наприклад, українське словникарство охоплює практику укладання словників українською мовою та усі видані/створені україномовні словники, а також власне видавництва та усіх осіб, і причетних до створення та реалізації словників в Україні.

З історичного боку словникарство є давнішим, ніж лексикографія, адже спочатку виникли короткі словники (ранні зібрання глос), а вже потім, набагато пізніше, виникла така галузь мовознавства як *лексикографія*. Перші європейські словникові праці постали ще у часи Античної Греції та Стародавнього Риму, а от власне (сучасна) європейська лексикографія (до якої зараховують й українську) зародилася лише у добу Відродження. Першим словником у європейській лексикографічній традиції є двомовний перекладний словник «Vocabolista italiano-tedesco» виданий у 1477 р. у Венеції.

Історія власне української лексикографії починається з кінця XVI ст. з перекладного словника Лаврентія Зизанія «Лексисъ съ толкованіемъ словенскихъ словъ», що був опублікований у 1596 р. у Вільні [Демська 2010, с. 46-50]. Одними з перших відомих українських лексикографів є Л. Зизаній, П. Беринда, Є. Славинецький, М. Максимович, Є. Желехівський, М. Уманець, Б. Грінченко.

Одним з центральних термінів лексикографії є словник. Класичне визначення *словника* у лексикографії подає О. Демська у своєму навчальному посібнику «Вступ до лексикографії»: *Словник — це певним чином організоване зібрання певних мовних одиниць, зазвичай слів, і приписаних до них коментарів, що описують особливості їхньої структури і/або функціонування, походження, перекладу з однієї мови на іншу/інші* [Демська 2010, с. 24]. Як видно з цього визначення словники можуть бути абсолютно різними, спрямованими на різні теми, з різними описами та структурою, але спільним для всіх є мовні одиниці та певна інформація про них. Кожний словник має свою макроструктуру.

Макроструктурою словника, або *структурою словника*, називається усе те, що наповнює словник. Усе що входить до змісту словника є його макроструктурою. Цей термін є широким і охоплює все що є в словнику [Демська 2010, с. 24].

До макроструктури словника входить його *реєстр*. Традиційно термін «реєстр словника» трактують, як упорядкований згідно з визначеним критерієм

чи критеріями список словникових одиниць, що підлягають опису, поясненню, з’ясуванню, перекладу тощо [Демська 2010, с. 24]. Тобто реєстр словника становлять всі ті одиниці, які підлягають опису в словнику. Реєстр залежить від того, які одиниці описуються в словнику. Якщо це буде кореневий словник, то реєстр словника буде складатися з усіх описуваних коренів у словнику, якщо афіксальний, то з афіксів. Отже, реєстр словника можуть формувати будь-які одиниці мовного рівня.

Реєстр кожного словника має свій список *реєстрових одиниць*, які своєю чергою входять до мікроструктури словника та є її заголовковими одиницями. Якщо макроструктурою називається усе те, що входить до словника, то *мікроструктурою словника*, або *словниковою статтею* називають певний текст, що пояснює, з'ясовує, трансформує (у перекладному словникові) тощо заголовкову чи реєстрову, одиницю словника, задаючи її основні характеристики, залежно від типу словника [Демська 2010, с. 25].

Мікроструктура словника дуже залежить від його типу та мети для якої він створювався, але вона обов'язково має два елементи: ліву *заголовкову одиницю*, або пояснюване (ту, що описують, і яка є паралельно реєстровою одиницею) та праву, або пояснювальне (інтерпретацію, пояснення до заголовкової одиниці) [Демська 2010, с. 25]. Щодо одиниць мікроструктури словника, то залежно від мети реалізації словника мікроструктуру словника можуть становити такі одиниці як: дефініція, екземпліфікація, ремарка, транскрипція, синопсис.

Отже, основними термінами лексикографії є: словник, макро- та мікроструктура, реєстр словника, реєстрова та заголовкова одиниці, дефініція, екземпліфікація, синопсис, транскрипція.

1.2.2 Типологія словників

Типологія словників — це класифікація словників за їхніми основними диференційними ознаками. Зазвичай класифікація словників має цілу систему ознак та критеріїв, за якими здійснюється класифікація. Усталеної та єдиної

типології словників немає. Одним з перших, хто спробував сформулювати критерії типізації словників був академік Л. Щерба. Окрім класифікації Щерби існує ще багато класифікацій словників. Деякі науковці намагалися створити таку типологію словників, яка б якнайповніше та якнайточніше змогла б класифікувати словники. Так Б. Городецький створив свою класифікацію словників, яка містить двадцять типологічних ознак словників [Перебийніс, Сорокін 2009].

Якщо звертатися до вітчизняних класифікацій, що загально типологізують словники, то найповнішу загальну академічну типологію словників наводить О. Тараненко в авторитетному академічному виданні «Енциклопедія Українська мова» [Тараненко 2004, с. 610-612].

Отже, немає єдиної загальноприйнятої універсальної типології словників, проте є загальні класифікації, які ставлять собі за мету загально класифікувати словник та ті, що класифікують лише певний тип словників.

1.2.3 Електронний словник як об'єкт комп’ютерної лексикографії

Комп'ютерна лексикографія займається створенням електронних словників — укладених за допомогою комп'ютера на основі комп'ютерної бази даних і розрахованих на користувача людину, або таких, що були укладені автоматично і розраховані на подальше їх використання машиною як бази даних (автоматичні).

Попри те, що найпоширенішим і найусталенішим терміном у мовознавстві є «комп'ютерна лексикографія», існують й інші терміни на позначення цієї наукової галузі: «електронна лексикографія», «обчислювальна лексикографія», «кібернетична лексикографія», «машинна лексикографія», «автоматична лексикографія» [Зубань 2020, с. 7]. Доречність вживання деяких з цих термінів, а саме «електронна лексикографія», замість усталеного «комп'ютерна лексикографія» є переконливою, а проте ми у своїй науковій роботі будемо користуватися вже усталеним у мовознавстві терміном.

Ще однією проблемою комп'ютерної лексикографії є її невизначений статус у множині мовознавчих наук. Це питання є дискусійним і переважно науковці схиляються до визначення комп'ютерної лексикографії як розділу прикладної (комп’ютерної) лінгвістики, що ставить лексикографічні завдання (Чепик, Палкова, Карпіловська, Голубкова та ін.), хоча деякі науковці схиляються й до вужчого трактування цієї наукової галузі, як одного з розділів лексикографії (Демська, Селегей та інші) [Зубань 2020, с. 8].

У цій науковій роботі ми будемо застосовувати широке розуміння терміна «комп'ютерна лексикографія», як розділу прикладної (комп’ютерної) лінгвістики у такому визначенні: «Комп’ютерна лексикографія — це прикладна наукова дисципліна в мовознавстві, що вивчає методи, технологію й окремі прийоми використання комп’ютерної техніки в теорії і практиці укладання словників. Спеціальні програми — бази даних, комп’ютерні картотеки, програми оброблення текстів — дозволяють в автоматичному режимі формувати словникові статті, зберігати словникову інформацію й обробляти її» [Зубань 2020, с. 8-9].

Щодо основних завдань комп'ютерної лексикографії, то в широкому розумінні цього терміна О. Зубань визначає такі основні завдання цієї наукової галузі:

1) використання комп’ютера в укладанні традиційних паперових словників;

2) автоматична конвертація паперових словників у електронні лексикографічні системи;

3) автоматичне / автоматизоване укладання електронних (комп’ютерних) словників;

4) автоматичне / автоматизоване укладання автоматичних словників [Зубань 2020, с. 9].

Безсумнівно словники, що були укладені за допомогою комп'ютера не вписуються у загальну класифікацію словників і потребують додаткової

класифікації. Подібну класифікацію наводять автори навчального посібника «Традиційна та комп'ютерна лексикографія» В. Перебийніс та В. Сорокін [Перебийніс, Сорокін 2009, с. 15] (див. Додаток 1).

Отже, *комп'ютерна лексикографія* — це наукова галузь, що займається створенням комп'ютерних словників за допомогою комп'ютерів та застосовує спеціальні методи для їх створення. Комп'ютерна лексикографія не обмежується лише створенням словників орієнтованих на використання людиною, але й також займається створенням словників орієнтованих на використання машиною.

1.3 Бази даних та бази знань особових імен

Оскільки кінцевим результатом роботи є база знань — ресурс, що зберігає й опрацьовує структуровані відомості, — доцільно стисло схарактеризувати базові поняття баз даних і баз знань та їхні типи, а також оглянути наявні електронні ресурси й бази знань, у центрі яких стоїть особове ім'я.

1.3.1 Основні поняття баз даних та баз знань

Опис майбутньої бази знань особових імен доцільно розпочати з розмежування трьох пов'язаних понять — даних, інформації та знань.

Даними називають окремі, не інтерпретовані факти про об'єкти реального світу, подані у формі, придатній для зберігання та опрацювання, — числа, символи, рядки тексту тощо. Самі по собі дані позбавлені контексту: послідовність літер «Віталій» лишається звичайним рядком символів доти, доки невідомо, що це особове ім'я. Коли даним надають значення та контекст, вони стають інформацією, а впорядкована й узагальнена інформація, придатна для висновків, формує знання. Саме послідовність від даних до інформації, а з неї до знань — лежить в основі розрізнення баз даних і баз знань [Карпіловська 2006, с. 34].

За способом організації розрізняють структуровані, неструктуровані та напівструктуровані дані. Структуровані дані підпорядковані наперед заданій

схемі та впорядковані за чіткою моделлю (до прикладу, таблиці з фіксованими стовпцями). Неструктуровані дані такої моделі не мають — це суцільний текст, зображення чи аудіо. Проміжне місце посідають напівструктуровані дані: вони не підпорядковані жорсткій табличній схемі, проте містять службові позначки, що відділяють одні елементи від інших (формати XML, JSON, JSONL). Саме до напівструктурованих належать дані, у форматі яких укладено нашу базу знань.

Упорядковане зібрання взаємопов'язаних даних, призначене для тривалого зберігання та багаторазового використання, називають базою даних, а програмний засіб для її створення й опрацювання — системою керування базами даних (СКБД).

За моделлю даних бази даних поділяють на реляційні та нереляційні. Реляційні бази даних, запропоновані Е. Коддом [Codd 1970], зберігають дані у вигляді таблиць (відношень), пов'язаних між собою через ключі, опрацьовуються мовою запитів SQL і вимагають чітко визначеної схеми. Нереляційні бази даних (NoSQL) такої жорсткої схеми не передбачають і охоплюють кілька різновидів: документоорієнтовані (зберігають записи як документи у форматі на кшталт JSON), «ключ — значення», стовпцеві та графові. Окремо вирізняють постреляційні бази даних, які розширюють реляційну модель, дозволяючи зберігати в одному полі вкладені, неатомарні значення. Саме така організація даних, як буде показано далі, найкраще відповідає структурі словникової статті з її численними варіантами імені.

Якщо база даних оперує власне даними, то база знань зберігає й знання про предметну область — відомості про способи застосування мовних об'єктів і судження про них, на основі яких можна робити умовиводи. Сукупність таких відомостей і утворює базу знань [Карпіловська 2006, с. 34].

Отже, на відміну від звичайної бази даних, база знань не лише зберігає структуровані відомості, а й дає змогу виводити на їх основі нову, прямо не зазначену інформацію. Саме ця властивість визначає характер створюваного нами ресурсу: база знань особових імен формуватиме цілісне знання про

антропонім, що охоплює його походження, деривaційні зв'язки (для похідних імен), а також варіантність як самого імені, так і його транслітерації.

1.3.2 Наявні електронні ресурси та бази знань особових імен

Перш ніж перейти до створення власної бази знань, доцільно розглянути вже наявні електронні ресурси, що описують особове ім'я або працюють із ним. Огляд охоплює дві групи джерел: вітчизняні відкриті ресурси й застосунки, орієнтовані на українські імена, та закордонні академічні бази знань власних назв, які демонструють усталені підходи до їх структурування.

Серед українських джерел найповнішим за наповненням є ресурс «Власні імена людей» [Власні імена людей], що дозволяє шукати ім'я за алфавітним покажчиком, у переліку імен або через поле пошуку, а за відсутності імені пропонує його відповідники з інших словників. Доступ є безкоштовним онлайн; програмний інтерфейс (API) надається лише за прямим зверненням до розробників, а відкритого репозиторію ресурс не має.

До ресурсів довідкового характеру належить і сайт «Статистика імен» [Статистика імен], що подає тлумачення імені, статистику його вживання та іменини. Відомості про динаміку називання (від 2009 р.) ґрунтуються на даних Головного управління статистики у Львівській області. Ресурс не має ані API, ані відкритого репозиторію, а частина зовнішніх покликань (зокрема на «Вікіпедію» та ресурс «Малеча») не працює.

Окрему групу становлять переліки українських жіночих та чоловічих імен у відкритій енциклопедії «Вікіпедія»: для жіночих імен зазвичай наведено варіанти, походження, значення та відомих носіїв, тоді як перелік чоловічих здебільшого обмежується самими іменами. Якість опису нерівномірна — не всі імена мають окремі статті, частина варіантів не збігається з авторитетним академічним словником «Власні імена людей», а серед джерел, на які покликаються автори, цього словника немає.

Іншу групу утворюють відкриті програмні засоби, орієнтовані радше на розробників. Бібліотека ***translit-ua*** [Chaplynskyi] здійснює транслітерацію української та російської мов за великою кількістю стандартів, зокрема

офіційним, і підтримує передавання українських слів засобами англійської, французької та німецької графіки; код є відкритим (ліцензія ***MIT*** [Open Source Initiative]), а автор покликається на відповідні стандарти (саме цю бібліотеку надалі використано для укладання бази даних транслітерацій). Вебзастосунок SpellingUkraine [Holub] призначений для перевірки правильного написання українських антропонімів та їх транслітерації; для нашого дослідження цінним є його підкорпус власних імен людей, що містить нормативне написання в оригіналі й транслітерації, а подеколи й значення імені. Бібліотека UA-morpher-toolkit [Lytvynenko] відмінює українські слова, у тому числі оніми, та визначає рід за іменем; вона так само має відкритий код і ліцензію ***MIT***.

Як бачимо, жоден з українських ресурсів не поєднує в межах єдиного машиночитного джерела повної інформації про особове ім'я — його варіантів, походження та транслітерації водночас: довідкові ресурси зосереджені на значенні та статистиці, а програмні засоби — на окремих операціях (транслітерації, відмінюванні).

Закордонний досвід репрезентують академічні бази знань власних назв із розвиненими моделями опису. Французький ресурс Prolexbase — багатомовний словник власних назв, що зіставляє імена та їхні варіанти між мовами. В його основу покладено мовнонезалежну типологію (4 супертипи та 34 типи), яка дає змогу впорядковувати синоніми й мовні варіанти; масив із десятків тисяч концептів сформовано з відкритих джерел, зокрема «Вікіпедії» та GeoNames [Maurel 2008]. Показовим є й досвід цифровізації етимологічного словника географічних назв SENGŚ, для якого розробники паралельно спроєктували реляційну модель та модель семантичного графа (RDF): саме граф дав змогу зберегти складні етимологічні описи, опрацювати офіційні й неофіційні варіанти та відобразити різний ступінь вірогідності історичних версій походження назви, не порушуючи жорсткої схеми [Tomasz 2021].

Подібний підхід реалізовано й в інших проєктах. Фінська інфраструктура відкритих пов'язаних даних NameSampo оцифрувала корпус із 2,7 млн географічних назв і опублікувала їх як граф знань, поділивши назви на мовні

складники, що уможливлює дослідження тенденцій називання у фінській, шведській та саамській мовах [Ikkala et al. 2018]. Оксфордський Lexicon of Greek Personal Names (LGPN) фіксує близько 400 тис. носіїв давньогрецьких імен і застосовує сутнісно-орієнтовану структуру запису, а його лінгвістичне розширення LGPN-Ling окремо аналізує морфологічну будову та етимологію кожного імені, відмежовуючи абстрактні мовні корені від конкретних історичних осіб [Lexicon of Greek Personal Names University of Oxford].

Отже, наявні українські ресурси описують особове ім'я лише частково, а закордонні бази знань, хоч і пропонують розвинені моделі, здебільшого не стосуються українського антропонімікону. Це засвідчує потребу в окремому машиночитному ресурсі, який інтегрував би відомості про українське особове ім'я, його варіанти, походження та транслітерацію, — саме таку базу знань створено в нашій роботі.

Висновки до першого розділу

Перший розділ кваліфікаційної роботи присвячено теоретичному підґрунтю створення бази знань особових імен. Оскільки така база поєднує мовний об'єкт, лексикографічне джерело та спосіб організації відомостей, її розгляд охопив три ділянки теорії — ономастику й антропоніміку, лексикографію та комп'ютерну лексикографію, а також теорію баз даних і баз знань.

Насамперед особове ім'я схарактеризовано в українському онімному просторі. Встановлено, що оніми вивчає ономастика, а особові імена — антропоніміка; з-поміж наявних класифікацій антропонімів для роботи обрано ту, що подана в енциклопедії «Українська мова». Розглянуто поняття особового імені та його варіантності.

Окрему увагу приділено транслітерації: подано визначення терміна та згадано офіційний стандарт транслітерації та показано, що поза межами паспорта імена транслітерують непослідовно, що породжує проблему

увідповіднення власних імен людей. Відповідно обґрунтовано доцільність додавання транслітерації до майбутньої бази знань.

Далі електронні словники та бази знань розглянуто крізь призму лексикографії. Розмежовано поняття лексикографії та словникарства й упорядковано базову термінологію. З'ясовано, що єдиної типології словників не існує, а найповнішою з вітчизняних є академічна типологія О. Тараненка. Окреслено й поняття комп'ютерної лексикографії, яку в роботі потрактовано широко.

Нарешті розглянуто теорію баз даних і баз знань. Розмежовано поняття даних, інформації та знань, а також типи даних і типи баз даних. Огляд наявних ресурсів засвідчив, що жодне з українських джерел не поєднує в межах єдиного машиночитного ресурсу відомостей про ім'я, його варіанти, походження та транслітерацію, а розвинені закордонні бази знань власних назв українського антропонімікону не охоплюють.

Отже, у першому розділі сформовано теоретичне та джерельне підґрунтя дослідження: визначено об'єкт опису, окреслено лексикографічний апарат і поняття баз даних та баз знань, а також обґрунтовано потребу в окремій базі знань українських особових імен, яку буде створено у практичній частині роботи.

РОЗДІЛ 2. СЛОВНИК-ДОВІДНИК «ВЛАСНІ ІМЕНА ЛЮДЕЙ» ЯК ДЖЕРЕЛЬНА ОСНОВА ДОСЛІДЖЕННЯ

Другий розділ присвячено джерельній базі дослідження — словнику-довіднику «Власні імена людей» Л. Скрипник та Н. Дзятківської, на основі якого укладатимемо базу знань. Оскільки якість майбутнього ресурсу безпосередньо залежить від обраного джерела, спершу обґрунтовано вибір саме цього словника з-поміж наявних антропонімічних праць, а далі проаналізовано його макроструктуру і мікроструктуру першого розділу, що безпосередньо слугуватиме матеріалом для наповнення бази знань. Кожному з цих питань присвячено окремий підрозділ.

2.1. Обґрунтування вибору словника-довідника «Власні імена людей» (Л. Скрипник, Н. Дзятківська, третє видання) як основного ресурсу для бази знань українських власних імен людей

Першочерговою метою створення бази знань українських власних імен людей є створення ресурсу, на основі якого можна проводити наукові дослідження, останнє неможливо забезпечити без використання словника авторитетного наукового видання. Таким джерелом є, зокрема, третє видання словника-довідника «Власні імена людей», підготовленого Л. Скрипник та Н. Дзятківською. У ньому особове ім'я схарактеризовано всебічно: словникова стаття охоплює широкий спектр відомостей — від офіційних і розмовно-побутових варіантів до докладного етимологічного коментаря та ілюстрації вживання. Зазначимо, що цей словник не є винятковою працею в українській антропонімії, відомими схожими працями є «Слов'янські особові імена українців: історико-етимологічний словник» П. Чучки, «Словник власних імен людей» С. Левченка, Л. Скрипник, Н. Дзятківської, «Словник українських імен» І. Трійняка та «Словник варіантів власних імен північно-західної України» Г. Аркушина.

З перелічених антропонімічних праць лише дві є у відкритому доступі: словник-довідник «Власні імена людей» та «Словник власних імен людей». В

останньому до реєстрової одиниці наведено лише її офіційний варіант, утворене від нього ім’я по батькові та його російський відповідник. Повнота опису особового імені та доступність словника-довідника «Власні імена людей» стали визначальними характеристиками для вибору його у ролі джерельної бази.

2.2 Макроструктура словника-довідника «Власні імена людей»

Макроструктура словника складається зі вступу, трьох розділів та додатків. У вступі Л. Скрипник коротко описує типи особових імен за походженням, наводить популярні та рідковживані імена, а також закликає батьків ставитися до вибору імені з усією серйозністю. Особливу увагу співавторка приділяє проблемі варіантності імен та нерозрізненню розмовно-побутових імен з документальними іменами, що часто призводить до непорозумінь та плутанини у документах. Для завчасного розв'язання таких проблем, а також для полегшення вибору імені дитини та надання ресурсу для мовознавчих досліджень і створювався цей словник-довідник [Скрипник, Дзятківська 2005, с.7-19].

Реєстровими одиницями у словнику виступають антропоніми, що були найбільш поширеними в офіційному вжитку на момент створення словника. Окрім онімів, що використовуються для офіційного найменування, у словнику також представлено розмовно-побутовий фонд українських імен. Хоча словник відображує великий пласт імен, що використовуються на території України, до його складу не увійшли діалектні варіанти імен, а також ті, варіанти імен, що вважаються зневажливо-згрубілими [Скрипник, Дзятківська 2005 с.20].

Словник виконує нормативну функцію, тому у додатках автори представили короткий ознайомчий список літератури з ономастики, а третій розділ словника присвятили правилам відмінювання та правопису власних імен людей в українській мові [Скрипник, Дзятківська 2005 с.28-29].

У другому розділі представлений двомовний словник особових імен людей, де перша частина є українсько-російською, а інша російсько-українською [Скрипник, Дзятківська 2005 с.25-28]. Що ж до першого розділу словника, то він

присвячений найпоширенішим особовим іменам в Україні. У цьому розділі біля кожного особового імені, що використовується в офіційному вжитку, представлені його повноправні відповідники, етимологічна довідка, а також його варіанти, що використовуються у розмовно-побутовому вжитку [Скрипник, Дзятківська 2005, с.20-25]. Оскільки саме цей розділ буде використаний для наповнення бази знань, його структуру надалі буде розглянуто детальніше.

2.3. Мікроструктура першого розділу «Найпоширеніші імена, їх походження та варіанти»

Перший розділ словника складається з трьох частин: «Чоловічі імена», «Жіночі імена» та «Покажчик скорочених та здрібніло-пестливих варіантів імен». Реєстрові одиниці у ньому розміщені за алфавітом, а власне словникова стаття складається з таких зон:

1. Повне офіційне ім’я
2. Повноправні нормативні варіанти імені
3. Варіанти імені, що трапляються у документах як повні, проте мають певні функціонально-стилістичні особливості (у дужках)
4. Мова-джерело
5. Коментар, щодо етимологічного походження оніма
6. Найбільш загальновживані скорочені та здрібніло-пестливі розмовно-побутові варіанти імені, що не увійшли до реєстрового ряду
7. Приклад вживання оніма [Скрипник, Дзятківська 2005 с.20-25]

Словникова стаття імені ***Євдокія*** чітко ілюструє цю структуру: *Євдокія, Докія (Явдоха, Вівдя) гр.; eudokia — благовоління; добра слава. ЄВДОСЯ; ДОКІЙКА, ДОКЄНЬКА, ДОКІЄЧКА, ДОСЯ, ДОСЕНЬКА, ДОСЕЧКА, ДОСЬКА, ДОЦЯ, ДОЦЕНЬКА, ДОЦЕЧКА, ДОЦЬКА, ДУСЯ, ДУСЕНЬКА, ДУСЕЧКА, ДУНЯ, ДУНЕНЬКА, ДУНЕЧКА; ЯВДОНЯ, ЯВДОНЕНЬКА, ЯВДОНЕЧКА, ЯВДОНЬКА, ЯВДОШЕНЬКА, ЯВДОШЕЧКА, ЯВДОШКА.*

— Се ви, Євдокіє? Здорові! — озвався Корній (Леся Українка);

Не кожна словникова стаття містить усі перелічені компоненти. Заголовкові одиниці іноді можуть не мати повноправних нормативних варіантів, наприклад: *ЗОЯ гр.; zoe — життя. ЗОЄНЬКА, ЗОЄЧКА, ЗОЙКА.* Подекуди заголовкові одиниці супроводжуються лише етимологічним коментарем та варіантом імені, що має певні стилістичні особливості, або лише етимологічним коментарем: *ФЕОДУЛ (ФЕДУЛ) гр.; theos — Бог і dulos — раб (буквально: раб Божий), СИЛЬВЕСТР лат.; silvester — лісовий.* Етимологічний коментар та екземпліфікація присутня майже у кожній словниковій статті, окрім тих, що посилаються на інші словникові статті.

Метапосилання є важливими складниками першого розділу, вони зменшують кількість повторення інформації. Всього таких метапосилань використовується чотири:

1. див.
2. те саме що
3. жін. до
4. утв. від чол. імені

Метапосилання ***див.*** та ***те саме що***, відрізняються тим, що перше посилається на словникову статтю з рівноправним йому варіантом імені, а друге посилається на варіант імені, що з часом став самостійним документальним іменем. Останні два: ***жін. до*** та ***утв. від чол. імені,*** посилаються на словникову статтю з чоловічим іменем. Відрізняються вони тим, що перше є співвідносним до чоловічого імені, а друге утворене від нього суфіксальним способом творення.

Окрім метапосилань у розділі також активно використовуються скорочення, що є спільними для всього словника. Вони зазвичай використовуються в етимологічному коментарі, аби вказати на походження імені, або уточнити іншу інформацію щодо нього [Скрипник, Дзятківська 2005, с.31].

«Покажчик скорочених та здрібніло-пестливих варіантів імен» є останньою з трьох частин першого розділу словника. Ця частина також поділена на частину з чоловічими та жіночими іменами. У кожній словниковій статті

цього підрозділу розмовно-побутові варіанти імен співвіднесено з офіційними документальними іменами та їх нормативними чи функціонально-стилістично обмеженими варіантами [Скрипник, Дзятківська 2005, с. 24]. До прикладу, так виглядає словникова стаття одного з чоловічих імен: *Жора, Жорж, Жорик — Георгій; Єгор; Юрій; Юрко.*

Оскільки саме перша частина першого розділу є найважливішою для створення бази знань, слід заздалегідь виокремити певні проблемні моменти, що у майбутньому можуть виникнути під час парсингу цієї частини.

По-перше, варто звернути увагу на варіативність словникової статті цього розділу. Загалом словникова стаття складається з 7 частин, кожна з яких, окрім заголовкової одиниці, може бути пропущена. Щонайменше словникова стаття може складатися з заголовкової одиниці, метапосилання ***див.*** та особового імені, що є офіційним варіантом до заголовкової одиниці. Щонайбільше словникова стаття може містити усі 7 частин, що у майбутньому може становити проблему, адже для того, аби розпарсити словникову статтю потрібно використовувати регулярні вирази, які вимагають чітких шаблонів. Відповідно, якщо не передбачити усі комбінації, існує ризик втратити частину словникових статей.

По-друге, почасти у словнику крапки літери «ї» зливаються, відповідно існує ймовірність, що при конвертації словника у текстовий формат ця літера буде замінена іншою.

Зрештою, проблема помилок у написанні особових імен є суттєвою, оскільки саме викривлення даних становить основний ризик — усі інші потенційні проблеми прямо чи опосередковано пов'язані з ним. Варто також зазначити, що в аналізованому розділі широко використовуються метапосилання. Якщо особове ім'я, на яке вказує таке посилання, відтворено з помилкою, пошук відповідної словникової статті стає технічно неможливим, що безпосередньо порушує цілісність і навігаційну зв'язність бази знань.

Висновки до другого розділу

Другий розділ кваліфікаційної роботи присвячений детальному обґрунтуванню та аналізу джерельної основи для створення бази знань особових імен людей. У результаті проведеного дослідження було визначено, що словник-довідник Л. Скрипник та Н. Дзятківської «Власні імена людей» (третє видання) є найбільш репрезентативним і вичерпним ресурсом серед наявних антропонімічних праць. Ключовими критеріями вибору стали його доступність та повнота опису, оскільки він надає комплексну характеристику особового імені, включаючи офіційні, розмовно-побутові варіанти, походження та ґрунтовний етимологічний коментар.

Було здійснено докладний аналіз макроструктури словника-довідника, яка охоплює вступ, три розділи та додатки, а також детально розглянуто його реєстр, що містить поширені офіційні антропоніми та розмовно-побутовий фонд особових імен. Основну увагу було зосереджено на мікроструктурі першого розділу «Найпоширеніші імена, їх походження та варіанти», який безпосередньо слугуватиме джерелом для наповнення бази знань. Встановлено, що словникова стаття цього розділу є варіативною і може містити до семи компонентів, зокрема повне офіційне ім’я, повноправні та функціонально-стилістично обмежений варіанти, етимологічний коментар, а також метапосилання (*див.*, *те саме що*, *жін. до*, *утв. від чол. імені*), які забезпечують навігаційну зв’язність.

За результатами структурно-функціонального аналізу було виявлено низку критичних проблемних моментів, які потенційно можуть ускладнити етап автоматичного парсингу. Серед ключових викликів виділено значну варіативність структури словникових статей, що ускладнює розробку універсальних регулярних виразів та підвищує ризик втрати даних. Окрім того, акцентовано на технічних ризиках, пов’язаних із конвертацією словника у текстовий формат, зокрема можливе викривлення літер української абетки та помилки в написанні особових імен. Особлива увага зосереджена на тому, що будь-які помилки у написанні імен можуть порушити цілісність метапосилань,

що є суттєвим ризиком для наповнення та коректності даних у майбутній базі знань.

У другому розділі було визначено джерельну базу дослідження, проаналізовано її макро- та мікроструктуру, а також виявлено потенційні технічні труднощі, які необхідно врахувати на наступному етапі — при практичній реалізації бази знань.

РОЗДІЛ 3. СТВОРЕННЯ БАЗИ ЗНАНЬ ОСОБОВИХ ІМЕН ЛЮДЕЙ

Третій розділ присвячено практичній реалізації бази знань особових імен на основі словника-довідника «Власні імена людей», аналіз якого було здійснено в попередньому розділі. Оскільки словник доступний лише у форматі PDF, тобто як комп'ютерна копія друкованого видання, процес створення бази знань охоплює кілька послідовних етапів: конвертацію словника у текстовий формат, парсинг словникових статей, розроблення схеми бази знань, формування бази даних особових імен та бази даних транслітерацій і, нарешті, їх об'єднання в єдину базу знань. Кожному з цих етапів присвячено окремий підрозділ.

3.1. Конвертація словника в текстовий формат

Словник-довідник «Власні імена людей» було отримано у форматі PDF. Зазвичай, якщо PDF-файл містить шар із накладеним текстом, то отримання тексту не є складним. Комп’ютерна версія словника-довідника містила текстовий шар, проте якість його була достатньо низькою. Оскільки від етапу конвертації словника залежала якість отриманих даних після парсингу, видобування тексту за допомогою оптичного розпізнавання символів (OCR) було найкращим рішенням, для того, аби гарантувати найкращу якість отриманого тексту.

Існує чимало реалізацій оптичного розпізнавання символів. Усі вони залучають ряд пайплайнів таких як: розпізнавання тексту, порядок читання сторінок, розпізнавання структури документа та інше. Зазвичай для кожного з цих завдань натреновано окремо нейронну мережу. Іноді використовуються великі мовні моделі для покращення якості розпізнавання або для власне розпізнавання. Усе це, а подекуди й більше, імплементовано у бібліотеках або доступно за допомогою API. До найпопулярніших з таких рішень відносять: **Mistral Document AI** [Mistral AI], **Marker** [Marker] та **Paddle OCR** [PaddlePaddle]. Усі три рішення є мультимовними та застосовують спеціально натреновані моделі для OCR. Різниця між ними полягає у ліцензії, використовуваних моделях, необхідних обчислюваних потужностей для їх

використання та ціні. **Marker** та **Paddle OCR** мають відкритий код і є бібліотеками, проте вони відрізняються ліцензіями використання. **Marker** використовує ліцензію ***AI Pubs Open Rail-M license*** [AI Pubs Open RAIL-M License v1], що дозволяє наукові дослідження, особисте використання та безкоштовна для стартапів, що мають прибуток до 2 мільйонів доларів. Своєю чергою **Paddle OCR** доступна під ліцензією ***Apache License 2.0*** [Apache Software Foundation. Apache License]**,** що дозволяє використання коду для наукових, особистих та комерційних цілей. Що ж до **Mistral Document AI**, то це є комерційним рішенням OCR від компанії **Mistral.** Доступ до моделей OCR відбувається за допомогою API. Перевагою цього рішення над іншими є відсутність залежності від обчислювальних потужностей, адже всі обчислення відбуваються на серверах компанії **Mistral**.

Оскільки від якості результатів OCR залежали інші етапи, було вирішено спочатку протестувати ці три рішення, а потім видобувати текст за допомогою найякіснішого рішення. Загалом для того, щоб обрати найкраще рішення, усі три було протестовано на 15 різних сторінках зі словника. Загальна інформація про рішення та результати швидкості конвертації представлені у таблиці (див. табл. 3.1).

Найшвидшим рішенням передбачувано стало рішення від **Mistral**. Серед локальних рішень обидва інструменти — **Marker** і **PaddleOCR** — успішно виконали завдання на GPU (NVIDIA GeForce RTX 5090). На GPU **Marker** виявився вчетверо швидшим за **PaddleOCR**. Швидкість рішення не є гарантом якості, тому для тестування якості OCR було обрано сторінки зі звичайним текстом [Скрипник, Дзятківська 2005, с. 30], списком [Скрипник, Дзятківська 2005, с. 31], словниковими статтями для жіночих [Скрипник, Дзятківська 2005, с. 115-118, с. 191-192, с. 128] та чоловічих імен [Скрипник, Дзятківська 2005, с. 50-55].

Назва	**Автор**	**Відкритий код**	**Швидкість конвертації (1 сторінка за секунду)**		
			назва моделі	**API**	**GPU**
Mistral Document AI	Mistral	ні	mistral-ocr-2512	0.08	-
PaddleOCR	PaddlePaddle	так	PaddleOCR-VL-1.5	-	5, 41
Marker	Datalab	так	surya	-	1.5

Таблиця 3.1

Порівняння OCR-рішень за основними параметрами

Аналіз роботи **Mistral Document AI**, **Marker** та **Paddle OCR** показав, що найгірше для сканованого україномовного словника спрацював **Paddle OCR** (див. табл. 3.2).

Paddle OCR (бібліотека від **PaddlePaddle**) заміняла українську літеру і у верхньому регістрі на и (ГЛИКЕРИЙ — ГЛИКЕРІЙ, ГОРДИЙ — ГОРДІЙ, АДЕЛИНА — АДЕЛІНА, ИРИНА — ІРИНА), або ігнорувала літеру «і» (ГЛБ — ГЛІБ, ГРИГР — ГРИГІР), так само як і літеру «ї» (АДЕЛАЙДА — АДЕЛАЇДА, АЛА — АЇДА). Попри те, що **Paddle OCR** краще справлявся із опрацюванням літер у курсиві (зокрема літерою «д»), ніж **Marker**, остання, як і **Mistral OCR,** не мала проблем з «і» та «ї». Загалом якість **Marker** та **Mistral OCR** виявилася дуже високою, проте **Marker** при опрацюванні літер у курсиві, а також літер із діакритикою повертала складні для опрацювання послідовності (*$\mathcal{A}\partial sica$*, *∂ .- $\epsilon \epsilon p$* замість курсивної літери «д»). Відповідно найкраще показало себе рішення від **Mistral**, модель не повертала незрозумілих символів, гарно опрацьовувала літери української абетки та літери з діакритикою.

Назва рішення	**Обробка літер української абетки**	**Обробка літер у курсиві**	**Обробка літер з діакритикою**
Mistral Document AI	Не знайдено жодних помилок	Не знайдено жодних помилок	У виключних випадках може видаляти діакритику.
PaddleOCR	Заміняє І на И, ігнорує Ї	Почасти погано обробляє літери у курсиві	Почасти заміняє літери з діакритикою на інші, ігнорує наголос.
Marker	Не знайдено жодних помилок	Майже усі літери у курсиві погано обробляє	Майже завжди ігнорує літери з діакритикою або заміняє їх на інші.

Таблиця 3.2

Якість розпізнавання засобами OCR

Опрацювання документа за допомогою **Mistral API** можна виконувати двома способами: опрацьовувати повністю документ або поділити його на частини й одночасно опрацьовувати. Попри те, що у коді було імплементовано обидва рішення, після детального аналізу відгуків про обрану модель (***mistral-ocr-2512)*** було вирішено опрацьовувати **перший розділ словника «Найпоширеніші імена, їх походження та варіанти»** повністю. Отриманий конвертований файл було розділено на три текстові файли: файл зі скороченнями, файл із чоловічими іменами (перша частина розділу) та файл із жіночими іменами (друга частина розділу).

3.2. Парсинг мікроструктури розділу

Текст є неструктурованим типом даних, аби перетворити його у структурований тип даних часто застосовують регулярні вирази. Їх особливістю є те, що вони дозволяють знаходити шматки тексту, що відповідають певному шаблону. Проте, їх використання має свої недоліки, основний з яких це так звана жадібність, що може призвести як до пропуску відповідного рядка тексту, так і до неправильного результату. З огляду на це парсинг словникових статей першого розділу словника було розділено на два етапи: виокремлення словникових статей з файлу та парсинг виокремлених словникових статей.

Метою першого етапу було перетворення текстових файлів зі словниковими статтями на окремі JSONL-файли, у яких кожен рядок відповідав одній словниковій статті. Для реалізації цього завдання було розроблено клас ***macro_parser.py***, що містив метод ***parse_dict***, який на основі визначеного регулярного виразу повертав файл у форматі JSONL, де кожний рядок мав всього два атрибути: id — унікальний ідентифікатор словникової статті та entry — текст словникової статті.

Оскільки схема бази знань не передбачала збереження прикладів використання особових імен, застосований регулярний вираз виокремлював фрагмент тексту від заголовкової одиниці до початку розділу з екземпліфікацією включно.

Окремої уваги заслуговує проблема форматування вхідних даних. У друкованому словнику частина статті без прикладів легко розпізнається завдяки виділенню імен жирним шрифтом — це стосується всіх імен, крім тих, що вжиті безпосередньо в прикладах. Однак текст, отриманий після конвертації за допомогою Mistral OCR, попри загалом високу якість розпізнавання, здебільшого не зберігав такого форматування. Відсутність виділення жирним шрифтом у конвертованому тексті суттєво ускладнила побудову регулярного виразу та потребувала додаткового опрацювання. Попри це вдалося створити ефективний регулярний вираз і в результаті виокремити 385 словникових статей із жіночими іменами та 457 із чоловічими.

Другий етап парсингу першого розділу словника передбачав три підетапи: розділення рядка словникової статті на її частини, перевірку отриманих результатів та їх постопрацювання.

Парсинг словникової статті, тобто розділення рядка словникової статті на частини відповідно до її структури, передбачав:

1. створення базових регулярних виразів,
2. конструювання регулярних виразів для словникової статті на основі базових регулярних виразів,
3. створення словника із впорядкованими регулярними виразами для словникової статті.

Останній підетап був зумовлений власне проблемою «жадібності» регулярних виразів. Словник зі створеними регулярними виразами потім передавався на вхід методу **parse_dict_entries_multi** класу **micro_parser.py** . У результаті було отримано два JSONL файли для чоловічих та жіночих імен. Кожний JSON рядок обов'язково складався з атрибутів entry_id та name. До необов'язкових атрибутів входили: official_vars, half_official_vars, equivalent_name, equivalent_name, base_name, etymology, etymology_comment, unofficial_vars (див. табл. 3.3).

Назва атрибута	**Значення**
entry_id	унікальний ідентифікатор словникової статті, що була отримана в результаті парсингу
name	заголовкова одиниця
official_vars	офіційні варіанти імені, тобто ті варіанти імені, що використовуються як документальні
half_official_vars	варіанти імені, що почасти використовуються як документальні

equivalent_name	офіційний варіант імені
base_name	ім’я від якого утворене, цей атрибут присутній лише у файлі з жіночими іменами
etymology	мова походження
etymology_comment	коментар, щодо походження імені
unofficial_vars	варіанти імені, що використовуються лише в побуті та не є документальними

Таблиця 3.3

Атрибути JSON-запису словникової статті

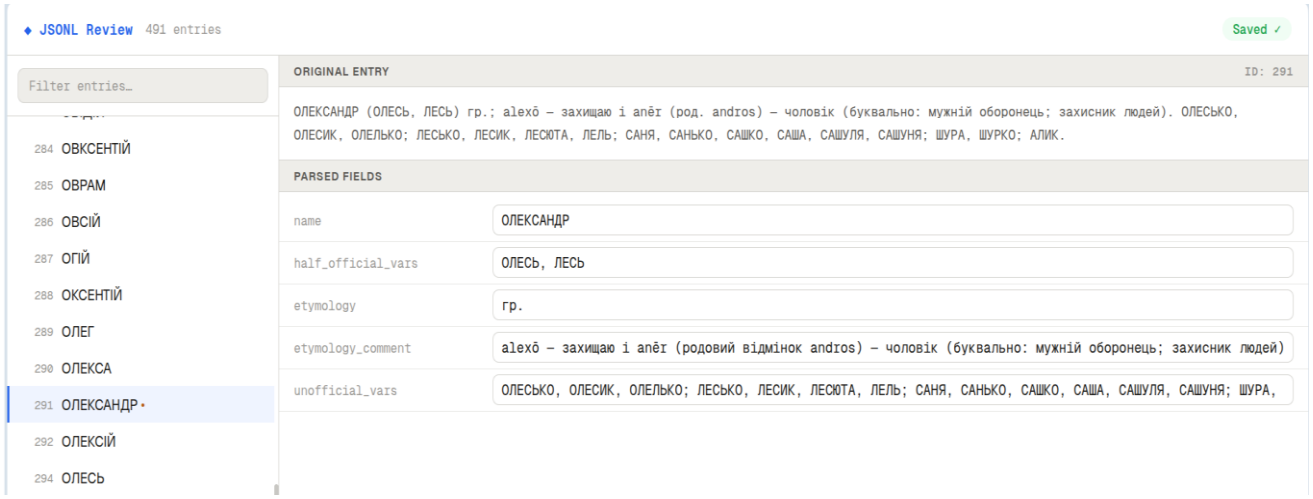
Важливо зазначити, що деякі атрибути є поєднанням властивостей, що розмежовувалися у попередньому розділі. Так було ототожнено метапосилання ***жін. до*** та ***утв. від чол. імені***. Таке рішення зумовлене кількома чинниками. По-перше, у словнику не зазначено, за допомогою якого суфікса жіноче ім'я утворено від чоловічого. По-друге, словникові статті, у яких жіноче ім'я співвідноситься з чоловічим, не мають етимологічного коментаря. Отже, окрім вказівки на чоловіче ім'я, такі метапосилання не містять жодної іншої інформації про особове ім'я. Тому обидва типи об'єднано, а наведені після них особові імена записано в атрибут ***base_name***. Подібне було зроблено із метапосиланням ***те саме що***, яке відсилалося на варіант імені, що з часом став документальним, особові імена після цього метапосилання записувалися в атрибут **official_vars**.

Наостанок варто зазначити, що у словнику зменшувально-пестливі варіанти, які набули статусу документальних, подаються окремо, натомість у нашій базі знань вони об'єднані з офіційними варіантами в атрибуті ***official_vars***.

Підетапи постопрацювання та перевірки отриманих результатів були майже одночасними.

Методологія перевірки результатів охоплювала створення спеціального інтерфейсу, аналіз помилок та їх виправлення. Щодо інтерфейсу, то кожну

заголовкову одиницю можна було побачити у лівій стороні екрана, а власне словникова стаття була розташована по центру. На ній можна було побачити рядок із повною словниковою статтею (результат першого етапу) та рядок JSON (результат етапу парсингу словникової статті) (див. мал. 3.1).



Малюнок 3.1

Інтерфейс програми для перевірки результатів парсингу

Етап перевірки результатів та аналіз помилок розкрив такі проблеми:

1. Попри те, що Mistral OCR на етапі тестування не мала проблем з українськомовним текстом, на практиці все ж було кілька постійних помилок. По-перше, використання І замість И в деяких іменах, наприклад: *ГАЛІНА* замість *ГАЛИНА*, *ЮХИМІНА* замість *ЮХИМИНА*. По-друге, у тексті замість літери Е іноді з'являлася російська літера Ё, наприклад: *ЕСТЁР* замість *ЕСТЕР*, *БОЛЁСЯ* замість *БОЛЕСЯ*, *СТЁПА* замість *СТЕПА*. На останок, іноді літеру Е заміняла літера Є, наприклад: *ЄВСЄВІЙ* замість *ЄВСЕВІЙ*, *ЄВСЄЙ* замість *ЄВСЕЙ*.
2. Словникові статті, що були погано скановані, іноді поверталися у комбінації латинки та кирилиці, наприклад: *ІРАЇDA* замість *ІРАЇДА*, *РАА* замість *РАЯ*.

3. Деякі імена містили дефіс у своєму складі, наприклад: *НАСТ-УСЯ*, *МАР-ІЙКА*
4. Проблему жадібності регулярних виразів вдалося мінімізувати, проте були випадки, коли опрацьована парсером словникова стаття містила у собі декілька словникових статей, наприклад: *МИЛА́Н слов.; від мил- (милий). МИ́ЛІЙ гр.; можливо, від Milyai — мілії (давнє плем'я, що населяло Лікію — країну на півд.-зах. узбережжі Малої Азії) або від mēlea — яблуна. МИЛОСЛА́В слов.; від мил- (милий) і слав- (слава). СЛА́ВА, СЛАВКО́, СЛА́ВЦБО, СЛА́ВИК, СЛА́ВЧИК. МИ́НА гр.; mēn — місяць (буквально: місячний). МИНКО́.*

Ці проблеми, окрім дефісів, були вирішені ручним редагуванням за допомогою інтерфейсу та редагування файлу. Дефіси, так само як і наголоси, були видалені автоматично за допомогою регулярних виразів. Також на етапі постопрацювання усі скорочення в атрибуті ***etymology*** було замінено його повним відповідниками, які були вказані у самому словнику.

У результаті цього етапу було отримано перевірені файли із словниковими статтями, що були опрацьовані засобами регулярних виразів. Цікаво, що кількість словникових статей збільшилася відносно, тої кількості, що була отримана на першому етапі, як для жіночих імен (409 на противагу 385), так і для чоловічих імен (498 на противагу 457).

3.3. Схема бази знань особових імен людей

Проєктування схеми бази знань є найважливішим етапом її створення. Попередньо в теоретичному розділі було визначено, що база знань реалізовуватиметься за допомогою постреляційної бази даних.

Центральною одиницею бази даних мало стати документальне особове ім'я. Таке рішення було прийнято з огляду на низку теоретичних і практичних аспектів.

По-перше, база даних мала містити інформацію про документальне особове ім'я, його офіційні документальні відповідники, варіанти імені, вживані лише в розмовному мовленні, а також його етимологію. Останню з практичного погляду значно легше пов'язати з одним документальним іменем, ніж з усіма його варіантами. По-друге, база даних мала укладатися на основі статей словника, опрацьованих методом парсингу, реєстровою одиницею якого є документальне ім'я, тож закономірно було залишити саме його в центрі опису.

Попри логічність обрання документального імені центральною одиницею, такий підхід означав би, що в базі даних мала б бути присутня транслітерація кожного варіанта імені — як офіційного, так і неофіційного. До того ж кожне ім'я мало б мати щонайменше дві форми транслітерації: офіційну та неофіційну. Такий обсяг інформації про одну-єдину одиницю зробив би базу даних надмірно громіздкою. З огляду на це було вирішено створити дві окремі бази даних: базу даних особових імен та базу даних транслітерацій особових імен. Кожна з них ґрунтуватиметься на результатах парсингу словникових статей та реалізовуватиметься в окремому файлі формату JSONL. Ключовою перевагою обраних постреляційних систем стане підтримка вкладених атрибутів, що дозволятимуть структурувати комплексні дані в межах одного об'єкта.

Відповідно до розробленої архітектури, кожна одиниця бази даних особових імен людей (JSON-об'єкт) базуватиметься на чотирьох основних атрибутах:

- ідентифікатор (***id***), що забезпечуватиме унікальність документального запису,
- власне документальне ім'я (***name***),
- описова характеристика (***description***),
- інформація про варіанти (***variant_info***).

Блоки ***description*** та ***variant_info*** проєктуватимуться як контейнери для вкладених параметрів. Так, атрибут ***description*** акумулюватиме дані про стать (***sex***), етимологію (***etymology***) та розширений етимологічний коментар (***etymology_comment***). Окремим необов'язковим компонентом тут виступатиме

base_name — він застосовуватиметься виключно для жіночих імен, що мають деривaційний зв'язок із відповідними чоловічими антропонімами. Своєю чергою, блок ***variant_info*** систематизуватиме списки офіційних (***official_vars***) та неофіційних (***unofficial_vars***) форм імені, що побутують у мовленні.

Що ж до бази даних транслітерацій особових імен, то вона міститиме чотири основні атрибути:

- ідентифікатор (id), що забезпечуватиме унікальність документального запису,
- власне документальне ім'я (name),
- офіційні варіанти імені (official_vars),
- неофіційні варіанти імені (unofficial_vars).

Кожне ім'я міститиме два вкладені атрибути — офіційну (***official_transliteration***) та неофіційну (***unofficial_transliteration***) транслітерацію. Атрибут ***official_transliteration*** зберігатиме форму імені, транслітеровану згідно з офіційним стандартом передавання українського алфавіту латиницею, затвердженим Постановою Кабінету Міністрів України № 55 від 27 січня 2010 року [КМУ 2010]. Натомість неофіційна транслітерація охоплюватиме всі інші системи, що наразі не мають офіційного статусу, проте колись були в ужитку.

Своєю чергою, атрибут ***unofficial_transliteration*** поділятиметься на три параметри — *uk_en*, *uk_gr* та *uk_fr*, кожен із яких міститиме список, елементами якого будуть транслітеровані форми імені. Важливо зазначити, що неофіційні транслітерації не виокремлюються в самостійні підтипи: поділ за мовами запроваджено винятково задля зручності, адже і німецька, і французька послуговуються розширеною латиницею з притаманною їм діакритикою.

Отже, у центрі бази знань особових імен буде документальне особове ім'я. Власне база знань складатиметься з двох баз даних: бази даних особових імен та бази даних транслітерацій особових імен. Обидві реалізовуватимуться засобами постреляційної бази даних у форматі JSONL.

3.4. База даних особових імен

Отримані на етапі парсингу першого розділу словника файли були структурованими та якісними, проте їх структура не повністю відповідала розробленій схемі бази даних. Між наявною структурою та цільовою схемою було виявлено кілька розбіжностей.

По-перше, дані потребували реорганізації у передбачені схемою контейнери ***description*** та ***variant_info***. По-друге, у вихідних файлах був відсутній атрибут для позначення статі (***sex***), який необхідно було додати. По-третє, записи подекуди мали атрибути ***half_official_vars*** та ***equivalent_name***, не передбачені схемою.

Окремої уваги потребувала відсутність даних про етимологію (***etymology***) та відповідних пояснювальних коментарів (***etymology_comment***) у частині записів. Ця проблема стосувалася насамперед статей, які містили атрибути ***equivalent_name*** або ***base_name*** і виконували в словнику роль метапосилань на інші реєстрові одиниці.

Для розв'язання цих завдань і приведення даних у відповідність до обраної схеми було розроблено модуль ***name_transformer.py***. Під час його створення постали два питання, що потребували вирішення: чи вилучати атрибут ***half_official_vars*** і як обробляти атрибут ***equivalent_name***.

З огляду на те, що видалення з опису будь-яких відомостей про документальне ім'я збіднило б його, інформацію, яка містилася в ***half_official_vars***, було вирішено не відкидати, а перенести до іншого атрибута. Відтак виникло питання, до якого типу варіантів — офіційних чи неофіційних — слід віднести ці імена. Оскільки ***half_official_vars*** становить перелік форм імені, що частково засвідчені як документальні, логічнішим видалося їх віднесення до атрибута ***official_vars***, тобто до офіційних варіантів документального імені.

Що ж до записів, які містили атрибут ***equivalent_name***, то було вирішено додавати всі офіційні варіанти імені на основі інформації зі словникової статті того імені, що зазначене в цьому атрибуті. Аналогічно було розв'язано проблему відсутності даних про етимологію (***etymology***) та відповідних пояснювальних

коментарів (***etymology_comment***): модуль запозичував ці відомості зі статей імен, зазначених в атрибутах ***equivalent_name*** або ***base_name***.

Після застосування модуля було отримано один файл формату JSONL із документальними жіночими та чоловічими іменами, загальна кількість яких становила 905 записів. Окремо варто зазначити, що було перевірено наявність атрибутів ***etymology*** та ***etymology_comment***. Перевірка показала, що, попри автоматичне додавання цих атрибутів для документальних імен, кілька записів усе ж не містили атрибута ***etymology***.

Таких записів виявилося сім, шість із яких — скорочені форми документальних імен, що з часом стали самостійними. Сьомий запис вирізнявся тим, що в самій словниковій статті мову-джерело прямо не вказано, а в коментарі зазначено, що науковці не дійшли спільної думки щодо походження цього імені, і наведено найпоширеніші версії. Після аналізу цих семи записів було вирішено залишити атрибут ***etymology*** порожнім — таке рішення дає змогу зберегти коректну структуру бази даних.

Отже, на цьому етапі файли формату JSONL з опрацьованими словниковими статтями було перетворено на базу даних особових імен, репрезентовану одним файлом формату JSONL обсягом 905 записів (див. файл **data/results/ukrainian_names_database.jsonl** на репозиторії [Mozghova]). За допомогою модуля ***name_transformer.py*** було сформовано структуру JSON-об'єкта відповідно до розробленої схеми: дані впорядковано в контейнери ***description*** та ***variant_info***, додано атрибут ***sex***, а відсутні значення ***etymology*** та ***etymology_comment*** заповнено на основі статей імен, зазначених в атрибутах ***equivalent_name*** і ***base_name***. Натомість не передбачені схемою атрибути ***equivalent_name*** та ***half_official_vars*** було вилучено, причому відомості з останнього атрибута було не відкинуто, а перенесено до атрибута ***official_vars***.

3.5. База даних транслітерацій особових імен

Формування бази даних транслітерацій особових імен передбачало насамперед вибір засобів, що дозволяли б якісно та якнайповніше транслітерувати ім’я. Існує кілька таких способів.

Перший — використати офіційний сайт Державної міграційної служби України [Державна міграційна служба України]. Він придатний для окремих осіб, які прагнуть дізнатися, як транслітерувати їхні ім'я та прізвище, проте непридатний для автоматичного оброблення сотень імен, оскільки сайт не надає публічного API.

Другий спосіб — розробити власний програмний код, що транслітеруватиме імена відповідно до чинного офіційного стандарту передавання українського алфавіту латиницею. Такий підхід дав би змогу швидко обробляти імена, однак потребував би розроблення та ретельного тестування власного алгоритму, а до того ж охоплював би лише чинний стандарт.

Оскільки база даних має містити не лише транслітерацію за чинним стандартом, а й за системами, що раніше були у вжитку, найдоцільнішим виявився третій спосіб — використання спеціалізованої бібліотеки. Після детального аналізу проєктів із відкритим кодом на платформі GitHub було обрано бібліотеку translit-ua [Chaplynskyi], створену українським дослідником Д. Чаплинським. Вона дає змогу транслітерувати україномовний текст за 13 транслітераційними таблицями, серед яких і чинний стандарт. Відповідно ця бібліотека лягла в основу модуля ***transliterator.py***, що відповідав за створення бази даних.

Результатом використання цього модуля стало 905 JSON об’єктів у JSONL файлі, кожний з яких відповідав визначеній схемі з одним незначним відхиленням: до списку *uk_en* потрапляла не кожна згенерована неофіційна транслітерація імені.

Аналіз отриманих транслітерацій показав, що деякі імена транслітеруються однаково за різними стандартами. Оскільки список *uk_en* об'єднує результати кількох систем транслітерації, орієнтованих на англійську

мову, а в межах неофіційної транслітерації не передбачено окремих атрибутів для позначення того, за яким стандартом утворено кожну форму, зберігати дублікати було недоцільно. З огляду на це було ухвалено рішення записувати до атрибута *uk_en* лише унікальні неофіційні транслітерації імені.

Отже, на цьому етапі було створено базу даних транслітерацій особових імен. Для її формування з-поміж розглянутих способів обрано спеціалізовану бібліотеку ***translit-ua***, що лягла в основу модуля ***transliterator.py***. Унаслідок його роботи було отримано 905 JSON-об'єктів, збережених в одному файлі формату JSONL (див. файл **data/results/transliterated_ukrainian_names_database.jsonl** на репозиторії [Mozghova]), кожен з яких відповідає розробленій схемі. Для кожного імені сформовано офіційну транслітерацію за чинним стандартом та неофіційні транслітерації за іншими системами, згруповані за мовами (*uk_en*, *uk_gr*, *uk_fr*), причому до списку *uk_en* записано лише унікальні форми.

3.6. База знань особових імен

На попередніх етапах було послідовно сформовано два окремі ресурси — базу даних особових імен та базу даних транслітерацій особових імен. Кожен із них реалізовано засобами постреляційної бази даних в окремому файлі формату JSONL і містить по 905 записів. Проте самі по собі ці файли є лише структурованими наборами даних; власне базою знань вони стають тоді, коли поєднуються в єдину систему, у центрі якої перебуває документальне особове ім'я, а між окремими записами встановлено зв'язки, що дають змогу не лише зберігати, а й виводити нову інформацію про ім'я.

Як було зазначено в теоретичному розділі, база знань відрізняється від бази даних насамперед наявністю зв'язків між одиницями та можливістю отримувати похідні відомості, які прямо не зафіксовані в окремому записі. У розробленій базі знань ця властивість реалізується кількома способами.

По-перше, кожен офіційний і неофіційний варіант співвіднесено з відповідним документальним іменем, тож, маючи на вході будь-яку розмовно-побутову форму, можна відновити її офіційний відповідник.

По-друге, на етапі трансформації (модуль **name_transformer.py**) відомості про етимологію та етимологічний коментар було успадковано з тих словникових статей, на які вказували метапосилання ***equivalent_name*** та ***base_name***; завдяки цьому навіть ті записи, що в самому словнику не мали власної етимологічної довідки, у базі знань отримали зведену інформацію про походження.

По-третє, для жіночих імен, утворених від чоловічих, у блоці ***description*** збережено атрибут ***base_name***, що фіксує деривaційний зв'язок між антропонімами (напр., *Августа* - це жіночий варіант до чоловічого імені *Август*).

Цілісність бази знань забезпечують спільний ідентифікатор (***id*** / ***entry_id***) та документальне ім'я (***name***), які виконують роль ключів, що пов'язують обидві бази даних. Запис з тим самим ідентифікатором в обох файлах описує те саме документальне ім'я: база даних особових імен подає його лінгвістичну характеристику (стать, етимологію, коментар, перелік офіційних та неофіційних варіантів), а база даних транслітерацій — латинізовані форми як самого імені, так і всіх його варіантів. Так, запис з ідентифікатором 1 в обох базах описує ім'я *Абакум*: у першій базі зафіксовано, що це чоловіче ім'я давньоєврейського походження з офіційним варіантом *Авакум* та низкою неофіційних форм (*Бакум*, *Абакумко* та ін.), а в другій — їхні транслітерації, зокрема офіційну *Abakum* і неофіційні *Abacum*, *Abakoum* тощо. Завдяки такому впорядкуванню за будь-якою формою імені користувач або програма можуть відновити повний «профіль» антропоніма.

Транслітераційний складник бази знань побудовано так, щоб охопити не лише чинну, а й історичні системи побуквеного передавання. Для кожного імені та кожного його варіанта сформовано офіційну транслітерацію за затвердженим стандартом, а також неофіційні транслітерації, згруповані за мовами орієнтації — *uk_en*, *uk_gr* та *uk_fr*. База даних транслітерацій так само налічує 905 записів — по одному на кожне документальне ім'я бази даних особових імен.

Створена база знань суттєво відрізняється від комп'ютерної копії словника-довідника, на основі якого її було укладено. По-перше, база знань містить інформацію, якої в самому словнику немає, — транслітерацію кожного

імені та його варіантів, чого не подає жодне з наявних джерел українських особових імен.

По-друге, метапосилання (***див.***, ***те саме що***, ***жін. до***, ***утв. від чол. імені***), які у словнику вимагають від читача самостійно переходити від однієї статті до іншої, у базі знань уже опрацьовано: пов'язані відомості (зокрема етимологію та офіційні варіанти) успадковано безпосередньо в записі, тож шукати їх в інших статтях не потрібно.

Водночас база знань не відтворює словник повністю: до неї не ввійшли приклади вживання імен (екземпліфікація), другий (двомовний) і третій (правописний) розділи словника та додатки — для наповнення використано лише перший розділ.

Наостанок, на відміну від словника, дані в базі знань нормалізовано: помилки розпізнавання виправлено, скорочення в позначенні мови-джерела розгорнуто до повних форм, а діакритику наголосу й дефіси, що іноді виникали при конвертації, вилучено.

У підсумку база знань охоплює 905 документальних особових імен, з-поміж яких 498 чоловічих та 407 жіночих; 131 жіноче ім'я пов'язане деривaційним зв'язком із відповідним чоловічим. Загалом до записів долучено 693 офіційні та 3545 неофіційних варіантів імен, кожен із яких має власну транслітерацію. За мовою-джерелом найбільше імен має грецьке (404), латинське (161), слов'янське (102) та давньоєврейське (101) походження; решта припадає на запозичені, новоутворені та інші. Лише поодинокі записи залишилися без зазначеної етимології — це переважно скорочені форми, що з часом стали самостійними іменами, та ім'я, щодо походження якого науковці не прийшли до спільної думки (як було описано в підрозділі 3.4).

Щодо атрибутів, то в базі даних особових імен обов'язкові атрибути — ідентифікатор (***entry_id***), документальне ім'я (***name***) та стать (***sex***) — заповнено для всіх 905 записів. Майже повністю заповнено й етимологічні атрибути: мову-джерело (***etymology***) зазначено у 897 записах, а етимологічний коментар (***etymology_comment***) — у 904. Решта атрибутів є необов'язковими і

заповнюються лише за наявності відповідних даних у словниковій статті: атрибут ***base_name*** наявний у 131 записі (виключно жіночі імена, похідні від чоловічих), офіційні варіанти (***official_vars***) — у 371, а неофіційні (***unofficial_vars***) — у 589 записах. У базі даних транслітерацій ідентифікатор (***id***) та документальне ім'я (***name***) заповнено для всіх 905 записів, причому кожна форма імені — як документальна, так і варіантна — обов'язково має офіційну (***official_transliteration***) та неофіційні (*uk_en*, *uk_gr*, *uk_fr*) транслітерації; блоки ***official_vars*** та ***unofficial_vars***, як і в базі імен, наявні лише там, де відповідне ім'я має такі варіанти.

Готову базу знань та код, за допомогою якого її було створено, було оприлюднено в публічному репозиторії на платформі GitHub [Mozghova], що відповідає одному із завдань дослідження — забезпечити її відкритість і доступність. Формат JSONL та постреляційна організація даних роблять базу знань придатною як для безпосереднього використання дослідниками й усіма, хто цікавиться українськими особовими іменами, так і для інтеграції в системи опрацювання природної мови, зокрема для завдань нормалізації та узгодження варіантів власних імен.

Отже, об'єднання бази даних особових імен та бази даних транслітерацій імен у єдину систему, пов'язану спільним ідентифікатором і документальним іменем, дало змогу створити базу знань українських особових імен. Вона не лише зберігає структуровані відомості про документальне ім'я — його стать, походження, офіційні та неофіційні варіанти і їх транслітерації, — а й дозволяє виводити похідну інформацію та переходити від будь-якої форми імені до її повного опису, що й відрізняє її від звичайної бази даних.

Висновки до третього розділу

Третій розділ кваліфікаційної роботи присвячено практичній реалізації дослідження — створенню бази знань українських особових імен на основі першого розділу словника-довідника Л. Скрипник та Н. Дзятківської «Власні імена людей». Оскільки джерельну базу було отримано у форматі PDF як

комп'ютерну копію словника, першим етапом стала конвертація тексту у придатний для опрацювання формат. З огляду на недостатню якість наявного текстового шару було застосовано оптичне розпізнавання символів (OCR). Порівняльне тестування трьох рішень — Mistral Document AI, Marker та PaddleOCR — на п'ятнадцяти сторінках словника засвідчило, що для українськомовного тексту з курсивом і діакритикою найкращу якість розпізнавання забезпечує Mistral Document AI, яке й було обрано для конвертації. Отриманий текст було поділено на три файли: зі скороченнями, чоловічими та жіночими іменами.

Наступним етапом став парсинг мікроструктури словникових статей, який, з огляду на «жадібність» регулярних виразів, було поділено на два кроки: виокремлення окремих словникових статей у формат JSONL (модуль ***macro_parser.py***) та власне розбір кожної статті на складові атрибути (модуль ***micro_parser.py***). Для перевірки й постопрацювання результатів було розроблено окремий інтерфейс, що уможливив виявлення та виправлення помилок, спричинених здебільшого недоліками розпізнавання (заміна літер І/И, Е/Ё, Е/Є, поєднання латиниці й кирилиці) та особливостями вхідних даних. Унаслідок цього етапу було отримано перевірені файли, що містили 498 чоловічих та 409 жіночих імен.

Ключовим етапом стало проєктування схеми бази знань. Було обґрунтовано рішення реалізувати її засобами постреляційної бази даних, поставивши в центр опису документальне особове ім'я. Оскільки зберігання транслітерацій усіх варіантів імені в межах одного запису зробило б базу даних надмірно громіздкою, було вирішено реалізувати базу знань у вигляді двох окремих баз даних — бази даних особових імен та бази даних транслітерацій особових імен, — кожну в окремому файлі формату JSONL із чітко визначеними атрибутами та вкладеними структурами.

Відповідно до розробленої схеми було сформовано обидві бази даних. За допомогою модуля ***name_transformer.py*** вихідні дані парсингу було приведено у відповідність до схеми: упорядковано в контейнери ***description*** і ***variant_info***,

додано атрибут статі (***sex***), вилучено не передбачені схемою атрибути, а відсутні етимологічні відомості успадковано зі статей, на які вказували метапосилання; результатом став файл обсягом 905 записів. Базу даних транслітерацій укладено на основі бібліотеки ***translit-ua***, що дала змогу транслітерувати кожне ім'я та його варіанти за чинним стандартом і за історичними системами, згрупованими за мовами; результатом так само стали 905 записів.

Поєднанням цих двох баз даних, пов'язаних спільним ідентифікатором і документальним іменем, було створено базу знань українських особових імен, яку оприлюднено в публічному репозиторії на платформі GitHub. На відміну від електронної копії словника-джерела, ця база знань є структурованим машиночитним ресурсом, що уможливлює автоматичний пошук та опрацювання даних, містить транслітерацію імен (відсутню в усіх наявних джерелах) і вже опрацьовані метапосилання, які не потребують ручного переходу між статтями.

Отже, у третьому розділі було досягнуто головної мети роботи — створено й верифіковано базу знань особових імен людей, придатну і для лінгвістичних досліджень, і для застосування в системах опрацювання природної мови.

ВИСНОВКИ

У кваліфікаційній роботі досягнуто поставленої мети — створено базу знань українських особових імен на основі словника-довідника Л. Скрипник та Н. Дзятківської «Власні імена людей». Реалізація мети передбачала послідовне розв'язання теоретичних і практичних завдань, узагальнені результати яких подано нижче.

На теоретичному етапі окреслено понятєвий апарат дослідження. Встановлено, що особове ім'я є предметом антропоніміки — розділу ономастики; з-поміж наявних класифікацій антропонімів обрано ту, що подана в авторитетній лінгвістичній енциклопедії, а саме поняття особового імені розглянуто разом із його варіантністю — співвідношенням офіційних і розмовно-побутових форм. Окремо схарактеризовано проблему транслітерації українського імені та чинний стандарт романізації, показано, що поза межами паспорта імена передають латиницею непослідовно: одне ім'я може мати до двадцяти одного варіанта, що породжує проблему увідповіднення власних імен людей. Опрацьовано також базову термінологію традиційної та комп'ютерної лексикографії й теорію баз даних і баз знань.

Здійснено огляд наявних джерел та застосунків, що репрезентують українські особові імена. Встановлено, що жоден із вітчизняних ресурсів не поєднує в межах єдиного машиночитного джерела відомостей про ім'я, його варіанти, походження та транслітерацію водночас, а розвинені закордонні бази знань українського антропонімікону не охоплюють. Це підтвердило потребу у створенні окремого ресурсу.

Обґрунтовано вибір джерельної бази: з-поміж антропонімічних праць найбільш репрезентативним і доступним визнано третє видання словника-довідника «Власні імена людей». Проаналізовано його макроструктуру та мікроструктуру першого розділу, що слугував матеріалом для наповнення бази знань. Встановлено, що словникова стаття є варіативною й може містити до семи компонентів, зокрема метапосилання, які забезпечують навігаційну зв'язність, а також виявлено потенційні труднощі автоматичного парсингу.

На практичному етапі реалізовано базу знань. Словник, доступний лише у форматі PDF, конвертовано в текст за допомогою оптичного розпізнавання символів (за результатами порівняння трьох рішень обрано Mistral Document AI). Здійснено двоетапний парсинг словникових статей із подальшою верифікацією результатів через спеціально розроблений інтерфейс. Спроєктовано схему бази знань на засадах постреляційної бази даних із документальним іменем у центрі опису та обґрунтовано її реалізацію у вигляді двох пов'язаних баз даних — особових імен і транслітерацій. Сформовано обидві бази даних (по 905 записів; базу транслітерацій укладено засобами бібліотеки ***translit-ua*** за чинним та історичними стандартами) і об'єднано їх у єдину базу знань, яку оприлюднено в публічному репозиторії на платформі GitHub.

Отже, усі заплановані завдання виконано, а мету досягнуто. На відміну від електронної копії словника-джерела, створена база знань є структурованим машиночитним ресурсом, що містить транслітерацію імен та вже опрацьовані метапосилання й уможливлює автоматичний пошук і опрацювання даних.

Створену базу знань можна використовувати в прикладних завданнях опрацювання природної мови, насамперед в увідповідненні власних імен людей, а також як джерело для антропонімічних досліджень і як основу для подальшого розширення — поповнення новими іменами, варіантами та додатковими відомостями про антропоніми.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Алексієнко, Л.А., Зубань, О.М. і Козленко, І.В., 2013. *Сучасна українська мова: морфологія*. За ред. А.К. Мойсієнка. Київ: Знання.
2. Андрієнко, Л.О., 2022. Поліваріантність особового імені як лінгвокультурний феномен. *Культура слова*, 96, с.228–238. https://doi.org/10.37919/0201-419X.2022.96.18.
3. Власні імена людей, н. д. *Словник власних імен людей*. [онлайн] Режим доступу: https://vlasni-imena-lyudej.slovaronline.com [Дата звернення: 26 Травня 2026].
4. Гонца, І.С., 2021. *Українська ономастика: навчальний посібник для здобувачів вищої освіти*. Умань: Візаві.
5. Демська, О.М., 2010. *Вступ до лексикографії: навчальний посібник*. Київ: Видавничо-поліграфічний центр «Київський університет».
6. Демчук, М.О., 1988. *Слов'янські автохтонні особові власні імена в побуті українців XIV–XVII ст.: монографія*. Київ: Наукова думка.
7. Державна міграційна служба України, н. д. *Перевірка транслітерації*. [онлайн] Режим доступу: https://dmsu.gov.ua/services/transliteration.html [Дата звернення: 26 Травня 2026].
8. Желєзняк, І.М., 2004. Антропонім. В: Русанівський, В.М., Тараненко, О.О. і Зяблюк, М.П. та ін. ред., 2004. *Українська мова: енциклопедія*. 2-ге вид., випр. і доп. Київ: Видавництво «Українська енциклопедія» ім. М.П. Бажана. с.29–30.
9. Зубань, О.М., 2020. *Комп'ютерне лексикографічне моделювання морфемної системи української мови: монографія*. Київ: Видавничо-поліграфічний центр «Київський університет».
10. Кабінет Міністрів України, 2010. Про впорядкування транслітерації українського алфавіту латиницею: Постанова № 55 від 27 січня 2010 р. [онлайн] Режим доступу: https://zakon.rada.gov.ua/laws/show/55-2010-п#Text [Дата звернення: 26 Травня 2026].

11. Карпенко, Ю.О., 2004. Власні назви. В: Русанівський, В.М., Тараненко, О.О. і Зяблюк, М.П. та ін. ред., 2004. *Українська мова: енциклопедія*. 2-ге вид., випр. і доп. Київ: Видавництво «Українська енциклопедія» ім. М.П. Бажана. с.83–84.
12. Карпіловська, Є.А., 2006. *Вступ до прикладної лінгвістики: комп'ютерна лінгвістика: підручник*. Донецьк: ТОВ «Юго-Восток, Лтд».
13. Кравченко, Л.О., 2002. *Антропонімія Лубенщини*. Автореф. дис. канд. філол. наук: спец. 10.02.01 «Українська мова». Київ: Київський національний університет імені Тараса Шевченка.
14. Перебийніс, В.І. і Сорокін, В.М., 2009. *Традиційна та комп'ютерна лексикографія: навчальний посібник*. Київ: Видавничий центр КНЛУ.
15. Редько, Ю.К., 1966. *Сучасні українські прізвища*. Київ: Наукова думка.
16. Русанівський, В.М., Тараненко, О.О. і Зяблюк, М.П. та ін. ред., 2004. *Українська мова: енциклопедія*. 2-ге вид., випр. і доп. Київ: Видавництво «Українська енциклопедія» ім. М.П. Бажана.
17. Скрипник, Л.Г. і Дзятківська, Н.П., 2005. *Власні імена людей: словник-довідник*. За ред. В.М. Русанівського. 3-тє вид., випр. Київ: Наукова думка.
18. Статистика імен, н. д. *База даних імен*. [онлайн] Режим доступу: http://database.ukrcensus.gov.ua/dw_name/main.asp [Дата звернення: 26 Травня 2026].
19. Тараненко, О.О., 2004. Словник. В: Русанівський, В.М., Тараненко, О.О. і Зяблюк, М.П. та ін. ред., 2004. *Українська мова: енциклопедія*. 2-ге вид., випр. і доп. Київ: Видавництво «Українська енциклопедія» ім. М.П. Бажана. с.610–612.
20. Трійняк, І.І., 2005. *Словник українських імен*. Київ: Довіра.
21. Худаш, М.Л., 1977. *З історії української антропонімії*. Відп. ред. Д.Г. Гринчишин. Київ: Наукова думка.
22. Чорноус, О.В., 2021. Антропоніми як базові складники онімної системи. *Записки з українського мовознавства*, 28, с.138–148. [онлайн] Режим

доступу: http://nbuv.gov.ua/UJRN/zukm_2021_28_16 [Дата звернення: 26 Травня 2026].

23. Чучка, П., 2005. *Прізвища закарпатських українців: історико-етимологічний словник*. За наук. ред. В. Німчука. Львів: Світ.
24. Чучка, П.П., 2011. *Слов'янські особові імена українців: історико-етимологічний словник*. Ужгород: Ліра.
25. Шульгач, В.П., 2016. *Нариси з праслов'янської антропонімії. Частина 3*. Київ: б. в.
26. AI Pubs, н. д. *Open RAIL-M License v1*. [онлайн] Режим доступу: https://www.licenses.ai/ai-pubs-open-railm-vz1 [Дата звернення: 26 Травня 2026].
27. Apache Software Foundation, н. д. *Apache License, Version 2.0*. [онлайн] Режим доступу: https://www.apache.org/licenses/LICENSE-2.0 [Дата звернення: 26 Травня 2026].
28. Chaplynskyi, D., н. д. *translit-ua*. [онлайн] Режим доступу: https://github.com/dchaplinsky/translit-ua [Дата звернення: 26 Травня 2026].
29. Christen, P., 2006. A comparison of personal name matching: techniques and practical issues. В: *Sixth IEEE International Conference on Data Mining — Workshops (ICDM'06)*, Hong Kong, China. Los Alamitos, CA: IEEE Computer Society. с.290–294. [онлайн] Режим доступу: https://users.cecs.anu.edu.au/~Peter.Christen/publications/mcd2006.pdf [Дата звернення: 26 Травня 2026].
30. Christen, P., 2012. *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Berlin: Springer. https://doi.org/10.1007/978-3-642-31164-2.
31. Codd, E.F., 1970. A relational model of data for large shared data banks. *Communications of the ACM*, 13(6), с.377–387. https://doi.org/10.1145/362384.362685.
32. Ecma International, 2017. *ECMA-404: The JSON data interchange syntax*. [pdf] 2nd edn. Geneva: Ecma International. Режим доступу: https://ecma-

international.org/wp-content/uploads/ECMA-404_2nd_edition_december_2017.pdf [Дата звернення: 26 Травня 2026].

33. Holub, O., н. д. *SpellingUkraine*. [онлайн] Режим доступу: https://github.com/Tyrrrz/SpellingUkraine [Дата звернення: 26 Травня 2026].
34. Ikkala, E., Tuominen, J., Raunamaa, J., Aalto, T., Ainiala, T., Uusitalo, H. and Hyvönen, E., 2018. «NameSampo»: a Linked Open Data infrastructure and workbench for toponomastic research. В: *Proceedings of the 2nd ACM SIGSPATIAL Workshop on Geospatial Humanities (GeoHumanities '18)*. New York: Association for Computing Machinery, Article 2. с.1–9. https://doi.org/10.1145/3282933.3282936.
35. Khairova, N. and Mozghova, A., 2025. Person name disambiguation in news articles: a hybrid method for enhancing entity resolution in Russia-Ukraine war coverage. В: *Proceedings of the Computational Linguistics Workshop at CoLInS 2025*. CEUR Workshop Proceedings, vol. 3976. с.43–54. [онлайн] Режим доступу: https://ceur-ws.org/Vol-3976/paper4.pdf [Дата звернення: 26 Травня 2026].
36. Lexicon of Greek Personal Names, н. д. *Lexicon of Greek Personal Names*. University of Oxford. [онлайн] Режим доступу: https://www.lgpn.ox.ac.uk [Дата звернення: 26 Травня 2026].
37. Lytvynenko, T., н. д. *UA-morpher-toolkit*. [онлайн] Режим доступу: https://github.com/Tymofii-Lytvynenko/UA-morpher-toolkit [Дата звернення: 26 Травня 2026].
38. Marker, н. д. *Marker*. [онлайн] Режим доступу: https://github.com/datalab-to/marker [Дата звернення: 26 Травня 2026].
39. Maurel, D., 2008. Prolexbase: a multilingual relational lexical database of proper names. В: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA).
40. Mistral AI, н. д. *Mistral Document AI*. [онлайн] Режим доступу: https://mistral.ai/solutions/document-ai [Дата звернення: 26 Травня 2026].

41. Mozghova, A., 2026. *Ukrainian-Names-Database* [Source code]. [онлайн] Режим доступу: https://github.com/anastasiia-mozg/Ukrainian-Names-Database [Дата звернення: 26 Травня 2026].
42. Open Source Initiative, н. д. *The MIT License*. [онлайн] Режим доступу: https://opensource.org/license/mit [Дата звернення: 26 Травня 2026].
43. PaddlePaddle, н. д. *PaddleOCR*. [онлайн] Режим доступу: https://github.com/PaddlePaddle/PaddleOCR [Дата звернення: 26 Травня 2026].
44. Patman, F. and Thompson, P., 2003. Names: a new frontier in text mining. В: *Intelligence and Security Informatics (ISI 2003)*. Lecture Notes in Computer Science, vol. 2665. Berlin: Springer. с.27–38.
45. SENGŚ, н. д. *Słownik etymologiczny nazw geograficznych Śląska*. [онлайн] Режим доступу: https://sengs.e-science.pl [Дата звернення: 26 Травня 2026].
46. Tomasz, K., 2021. Digital transformation of the etymological dictionary of geographical names. *Applied Sciences*, 11(1), 289. https://doi.org/10.3390/app11010289.
47. Ward, I., н. д. *JSON Lines*. [онлайн] Режим доступу: https://jsonlines.org/ [Дата звернення: 26 Травня 2026].

ДОДАТКИ

Додаток 1

Типологія комп'ютерних словників В. Перебийніс та В. Сорокіна