СТРУКТУРНА, ПРИКЛАДНА ТА МАТЕМАТИЧНА ЛІНГВІСТИКА

УДК 378:81
DOI https://doi.org/10.32838/2710-4656/2021.4-2/28

Гриців Н. М.
Інститут комп'ютерних наук та інформаційних технологій
Національного університету «Львівська політехніка»

Бабяк С. А.
Інститут комп'ютерних наук та інформаційних технологій
Національного університету «Львівська політехніка»

Софіяник Н. І.
Інститут комп'ютерних наук та інформаційних технологій
Національного університету «Львівська політехніка»

ПРИКЛАДНІ АСПЕКТИ ЛІНГВОТЕХНОЛОГІЙ

У статті проаналізовано прикладні аспекти лінгвотехнологій. Зокрема, розглянуто стрижневі лінгвотехнологічні принципи створення електронних термінологічних словників із покликанням на чужоземний досвід комп'ютерної лексикографії та запропоновано огляд можливостей автоматизації обробки природної мови.

З лексикологічного погляду вибір об'єкта наукового зацікавлення зумовлений підвищенням необхідності відображення понятійно-термінологічного апарату прикладних царин, позаяк понятєву систему піддають упорядкуванню, критичному й розлогому аналізу, чіткому окресленню, термінологічні словники слугують основними засобами нормалізації науково-технічної термінології. Зокрема, тлумачно-перекладні словники слугують особливим різновидом термінологічних.

*Тому в статті окреслено основоположні етапи генерування словника, як-от: визначення джерельної бази, автоматизація процесу відбору, попередня обробка, відбір термінів і статистичне ранжування. Окрім цього, виокремлено новітні процеси попередньої обробки матеріалу (напр., токенізація, лематизація, частиномовне розмічування), а також розглянуто передові рішення для автоматизації витягу термінологічних даних (напр., інноваційна хмарна платформа **TaaS**).*

*Окрім лексикографічних перспектив, прикладні аспекти лінгвотехнологій мають потенціал в обробці не окремих лексем, а цілих текстових масивів. Відомо, що протягом останніх двох століть людство ефективно використовувало автоматизацію численних завдань, застосовуючи механічні й електричні технології. Маючи це на гадці, в цій статті також розглядаємо проблему автоматизації обробки природної мови, а саме: кодування текстів за допомогою ініціативи **TEI (Text Encoding Initiative)**. Поряд подано загальну характеристику **TEI** як міжнародну спільноту дослідників природної мови.*

*Оскільки проблеми опрацювання природної мови далеко не однозначні, їхнє дослідження викликає істотне наукове зацікавлення. У статті також наведено узагальнену характеристику принципів системи **TEI**. Значну увагу приділено огляду різних наборів розмітки, а саме: **TEI P3** та **TEI P5**, їхнім спільним і відмінним рисам для опису структури, зовнішнього вигляду та змісту тексту.*

Ключові слова: *прикладна лінгвістика, природня мова, токенізація, лематизація, розмітка, термінологічна база даних, лінгвотехнології, стандарт TEI, ініціатива кодування тексту, тег.*

Постановка проблеми. Укладання словника слугує однією з найбільш трудомістких діяльностей людини в царині лексикографії. Відбір належного матеріалу, його впорядкування, «доставка» до кінцевого користувача потребують чимало ресурсів, як-от: час і людські зусилля. Позаяк паперові словники поступово втрачають свою актуальність, маємо намір зосередитися на вивченні засадничих лінгвотехнологічних принципів створення електронних термінологічних словників з огляду на чужоземний досвід.

Окрім цього, збереження текстів в електронному форматі надає багато нових можливостей, як-от: висока швидкість пошуку інформації, легкість редагування, мультимедіа, гіперпосилання. Здебільшого відсканові тексти можуть зберігатись у різних графічних формах. Такі формати певною мірою підходять для відтворення на великих дисплеях. Проте для ефективнішого застосування відсканові тексти можна перетворити на цифрові текстові формати. Як результат, потрібно розпізнати текст і перетворити інформацію із графічного формату на електронний текст. Цифровий формат тексту зменшує розмір файлу та дозволяє іншим програмам переформатувати, шукати або обробляти текст.

Аналіз останніх досліджень і публікацій. Ми покликаємось на погляди таких дослідників, як-от: М. Пінніс, Т. Ґорностай, Р. Скадіньс, А. Васільєвс. Ми детально розглянули їхню статтю *«Online platform for extracting, managing, and utilising multilingual terminology»* («Онлайн-платформа для витягу, управління та використання мультилінгвальної термінології»), оприлюднену у 2013 році [4]. Ми також звертаємось до ідей інших зарубіжних науковців, з-поміж яких – І. Рьозіґер, Йо. Шефер, Т. Джордж, С. Таннерт, У. Хайд, М. Дорна. Ми дослідили їхню статтю *«Extracting terms and their relations from German texts: NLP tools for the preparation of raw material for specialized e-dictionaries»* («Витяг термінів і їхніх зв'язків із німецьких текстів: *NLP*-інструментарій для підготовки первинних даних для спеціалізованих електронних словників»), опубліковану у 2015 році [6].

Багато зарубіжних розвідок, особливо дослідження таких дослідників, як Д. Журафськи, Дж. Х. Мартін чи Р. М. Різ, зосереджуються на вивченні обробки природних мов. Також дослідження В. Ю. Таранухи та Н. П. Дарчук стосуються обробки природних мов й автоматизованого синтаксичного аналізу в українській комп'ютерній лінгвістиці.

Постановка завдання. Ми націлені описати теоретичне підґрунтя генерування термінологічного словника, визначити стрижневі етапи й лінгвотехнологічні засоби його створення, а також прагнемо описати основні засади Ініціативи кодування тексту (*TEI*), їхні принципи та проаналізувати схеми кодування *TEI P3* та *TEI P5*.

Виклад основного матеріалу. Кожні два роки відбуваються *eLex*-конференції, на яких дослідники вносять новаторські пропозиції, спрямовані на стрімкий розвиток електронної лексикографії. Стаття *«Online platform for extracting, managing, and utilising multilingual terminology»* [4] покликана запропонувати передові рішення для автоматизації витягу термінологічних даних. Дослідники представляють інноваційну хмарну платформу *TaaS* (*Terminology as a Service*), яка надає користувачам доступ до термінологічних даних як витягу з завантажених документів, зіставних чи паралельних корпусів. Тож *TaaS* слугує засобом для підготовки статей моно- та білінгвальних словників.

Із метою підготовки матеріалу для монолінгвальних словників *TaaS* виконує певні кроки, як-от:

− витяг тексту з документів, завантажених користувачем;

− відбір термінів як наслідок частиномовної розмітки;

− оцінка термінів, відповідно до певних статистичних показників;

− вилучення всіх унікальних термінів із розмічених документів;

− нормалізація вилучених термінів.

Після виконання вищенаведених автоматизованих процесів користувач може провести «фільтрацію» термінів чи здійснити відповідний переклад, щоб перейти до створення мультилінгвальних словників. Однак платформа спроможна відібрати дані як матеріал і для білінгвальних термінологічних словників, позаяк *TaaS* здійснює пошук можливого перекладу для кожного з вилучених термінів. Для цього платформа використовує термінологічні бази даних, з-поміж яких – *EuroTermBank*, *IATE* (*InterActive Terminology for Europe*), *TAUS* (*Translation Automatic User Society*). Окрім цього, *TaaS* застосовує дані, узяті з паралельних і зіставних корпусів, з яких відразу вилучає необхідні терміни з відповідним перекладом. Платформа може послуговуватися й мультилінгвальними корпусами, що уможливлює збір даних для мультилінгвальних словників.

Розгляньмо й матеріали статті *«Extracting terms and their relations from German texts: NLP tools*

for the preparation of raw material for specialized e-dictionaries» [6]. Укладання словника – доволі клопіткий процес, який вимагає чіткого дотримання його основоположних етапів. Передусім необхідно **визначити джерельну базу**, з якої відбиратимемо необхідний матеріал.

Надалі варто **автоматизувати процес відбору** матеріалу. Дослідники ілюструють засоби обробки природної мови, методи й технології, використані для витягу термінологічних даних. Науковці послуговуються засобом для витягу термінів, що поєднує попередню обробку лінгвістичного корпусу зі статистичним ранжуванням. З-поміж етапів **попередньої обробки** виокремлюють:

– токенізацію, що полягає в розмежуванні й розмічуванні речень і словоформ;

– частиномовне розмічування й попередню лематизацію засобами *RF*-тегувальника;

– лематизацію як особливу обробку словоформ, відсутніх у лексиконі тегувальника.

Після попередньої обробки матеріалу визначають стадію, що полягає у **відборі термінів**. Задля цього застосовують певні моделі, що ґрунтуються на *POS*, тобто частиномовній розмітці. З-посеред типових іменникових термінологічних моделей виокремлюють іменники, сполуки прикметника й іменника, іменники, що вимагають модифікатора в родовому відмінку чи прийменника з відповідним модифікатором. Для витягу дієслівних термінів послуговуються парсером залежності, тренування якого здійснено на основі корпусу *TiGer Treebank*. Парсер виконує синтаксичну розмітку текстів.

Подальша стадія стосується **статистичного ранжування**, що полягає в сортуванні списків, сформованих на попередніх етапах, із вмістом відповідних термінів. Унаслідок сортування отримують належний список термінів.

Необхідно зазначити, що вищенаведені етапи (за підтримки зазначених методів і засобів) призначені для полегшення праці лексикографа, а саме: для підготовки словникових статей. Матеріал, зібраний вищеподаними засобами, не націлений на цілковиту автоматизацію створення лексикографічного продукту, а на представлення необхідних іменникових і дієслівних елементів.

А тепер розгляньмо стрижневі засади Ініціативи кодування тексту (*TEI*). Наявність сотень природних мов, кожна з яких має свій набір синтаксичних правил, ускладнюють опрацювання природної мови. У межах однієї мови слова можуть мати багато значень залежно від контексту. Навіть на рівні символів є певні ускладнення.

TEI є машиночитаною мовою, яка полегшує роботу з текстом, виділяючи важливу інформацію тегами. Крім того, з 1980 р. Ініціатива кодування тексту (*TEI*) є всесвітньою групою дослідників письмової мови, що працюють у сфері практики цифрових гуманітарних наук. Станом на сьогодні спільнота *TEI* веде список розсилки, серій зустрічей та конференцій, а також журнал, сховище *GitHub* та набір інструментів, а також підтримує єдиний технічний стандарт. З 2001 року спільнота *TEI* стає консорціумом, а саме: об'єднує зусилля фахівців у галузі комп'ютерної обробки природних мов для досягнення бажаного результату.

Основні принципи системи *TEI* визначають: (а) здатність ефективно здійснювати різні види досліджень; (б) простоту, чіткість і конкретність; (в) незалежність у використанні спеціального програмного забезпечення; (г) здатність точно ідентифікувати й ефективно обробляти текст; (ґ) перспективу користувацьких доповнень; (д) відповідність чинним і новим стандартам. Принципи системи *TEI* вперше опубліковані в 1994 році. Низка попередніх рекомендацій щодо принципів *TEI* з'явилася в 1990–92 рр.: 1990 (*P 1*), 1990 (*P 1.1*), 1992 (*P 2*). З 1994 по 2002 рр. їх опублікували як стандарти для кодування й обміну електронними текстами – *Guidelines for Electronic Text Encoding and Interchange*: 1994 (*P3*), 1999 (*P3.1*), 2002 (*P4*), які містять синтаксис (набір *SGML* або *XML*) і семантику (розмітку структури тексту, що визначається низкою тегів й електронним заголовком *TEI* у документації *TEI*) [1, с. 424].

У процесі обробці природної мови слід постійно враховувати кодування, яке використовують у цьому документі. Розділові знаки та цифри можуть вимагати особливого поводження. Часто потрібно аналізувати використання символів, що передають емоції (сполучення літер або специфічних символів), гіперпосилання, повторювані знаки пунктуації (... або ---), розширення файлів й імена користувачів, включаючи крапки окремо [5].

TEI, на відміну від інших форматів кодування, є специфічною мовою розмітки, яка містить електронні текстові джерела, інформацію про автора, вихідні дані, першоджерела, налаштування рукопису й іншу інформацію.

Стандарт *TEI* створено в 1987 р. та стандартизовано в 1990 р. Це спроба надати найбільш вичерпні інструменти для розмітки будь-якого тексту; він містить єдину систему, а також збірник керівних принципів і практик. На відміну від інших форматів, *TEI* можна вдосконалити й модифікувати для задоволення конкретних потреб [8].

Отже, *TEI* як ініціатива кодування тексту є світовим і мультидисциплінарним стандартом для зображення всіх типів текстів із використанням найбільш виразної та найменш застарілої схеми кодування. Обрана схема кодування текстів має відповідати загальній системі *TEI*. Дотримання цієї вказівки перевіряють принципи кодування й обмін електронними текстами (*Guidelines for Electronic Text Encoding and Interchange*). Принципи *TEI* створено з урахуванням різноманітних застосувань і дисциплін, що дозволяє їм обробляти величезні обсяги текстів, будучи при цьому максимально узагальненими й адаптованими.

TEI використовують для впорядкування тексту у вигляді «дерева». Весь документ вважається «кореневим елементом» із певними ознаками, такими як розділи, глави, сторінки, абзаци, заголовки тощо, що відгалужуються від кореня. Саме ця структура дає змогу шукати документ *TEI* й застосовувати таблиці стилів для кращого показу для користувача. *TEI* може бути налаштованим відповідно до потреб проєкту; крім цього, включає теги, специфічні для певного жанру – драма, поезія, проза.

Для початку роботи з *TEI*, текст, який вибраний для кодування, має бути перенесеним у текстовий редактор. Найбільш частими способами передачі слугують введення тексту вручну та сканування тексту за допомогою оптичного розпізнавання тексту. Збереження первинного формату тексту та друкарських характеристик сприяє відповідному розміщенню тегів *TEI*. Ці теги можна використовувати для визначення розривів, абзаців і рядків, а також основних частин тексту, таких як заголовки глав або розділів. Теги також можна розміщувати навколо друкарських елементів, таких як курсив або підкреслений текст, перенесення, спеціальні символи, як-от: знак амперсанда чи долара, а також альтернативні написання й орфографії.

Використання таблиці стилів або іншого інструменту перетворення необхідне для візуальних презентацій документа, кодованого *TEI*. Іншими словами, документи *SGML* або *XML*, звичайно, не призначені для читання «необробленими». Їх можна використовувати з програмним забезпеченням, яке зчитує теги як «поля» бази даних під час пошуку та як набір інструкцій типографського макета під час показу результатів.

Певний текст можна закодувати у *TEI P3* (*SGML*), оскільки це описова схема кодування, яка дозволяє досліднику пояснювати структуру та семантику текстових функцій, які він хоче проаналізувати. Наприклад, документ кодується в елемент <*TEI.2*>, що містить як розділ <*teiHeader*> для метаінформації та частину <*text*> для фактичного змісту тексту. Заголовок має містити мінімальну кількість метаінформації, тоді як сам текстовий вміст кодується в <*body*> [7]. Усередині тексту структурні елементи (заголовок – <*head*>, абзац – <*p*>, виноска – <*note @ place = = foot*>), а також семантичні ознаки (заголовок – <*title*>, виділення – <*emph*>, термін – <*term*>) можуть бути повністю вираженими за допомогою зрозумілих назв тегів [7].

Однак зауважимо, що в *SGML* деякі елементи можуть вживатися без кінцевих тегів (<*title*>, <*body*>, <*p*>, <*head*>), а значення атрибутів можуть бути без оточувальних лапок (“type = foot”).

Окрім цього, текст також можна закодувати в *TEI P5* (*XML*). Остання версія керівних принципів *TEI* визначає схему описового кодування у форматі *XML*. Як можна побачити, у *TEI P5* є багато подібностей із кодуванням *TEI P3*; наприклад, усі структурні та семантичні текстові функції можна позначити досить інтуїтивними назвами елементів. Проте наявні деякі відмінності: у *TEI P5* усі елементи повинні мати кінцеві теги; у *TEI P5* усі значення атрибутів повинні бути оточеними лапками; деякі імена основних елементів змінено (наприклад, перший елемент будь-якого тексту *TEI P5* тепер називається <*TEI*>) [7].

Загалом 88 проєктів були проведені або проводяться відповідно до принципів *TEI*. Паралельно з повною версією принципів *TEI*, яка визначає сотні елементів *SGML*, широко використовують документ *TEI Lite*, метою якого є надання так званого початкового набору елементів *SGML* для кодування тексту користувачем. *TEI Lite* зосереджує в собі ядро набору тегів *TEI*, обробляє широкий спектр текстів, дозволяє розробляти нові документи для кодування тексту, і, що є найсуттєвішим, він компактний і простий у використанні.

Висновки і пропозиції. Отже, ми дослідили прикладні аспекти лінгвотехнологій. Зокрема, ми виокремили стрижневі початкові етапи генерування словника: визначення джерельної бази, автоматизація процесу відбору, попередня обробка, відбір термінів і статистичне ранжування. Ми розглянули новітні методи, які покликані автоматизувати процес укладання словників, а саме: відбір матеріалу для словникових статей. Ми наголошуємо на застосуванні інноваційної хмарної платформи *TaaS*, яка уможливлює витяг даних для укладання моно-, бі- та мультилінгвальних

термінологічних словників. Токенізація, лематизація та частиномовне розмічування сприяють полегшенню відбору матеріалу на етапі його попередньої обробки. Як засвідчує чужоземний досвід, лексикографи послуговуються численними засобами, які полегшують їхню працю, а отже, сприяють швидкісному розвитку електронної лексикографії.

Також можемо зробити висновок, що основною метою *TEI* є забезпечення належного обміну текстовою інформацією в електронній формі. Крім того, *TEI* надає спосіб пояснити архітектуру тексту, щоб полегшити його обробку за допомогою програмного забезпечення, що працює на різних комп'ютерних системах і в різних технічних налаштуваннях. Отже, кодування первинних даних має бути організованим згідно з ідеєю *TEI* і пропонує кодування для загальної структури первинних даних, типографської розмітки та редакторських правок, одиниць рівня абзаців і рівня речень.

Наявність значної кількості несумісних систем кодування та збільшення сфери використання електронних текстів були передумовами для створення системи *TEI*. Стандарт *TEI* створює відповідну рівномірність між загальною концепцією подання природної мови та простотою реалізації кодування. *TEI* також пропонує різноманітні методи зображення лінгвальної та металінгвальної інформації.

Навіть з огляду на поверхневе вивчення основних принципів кодування даних *TEI* виявляє, що будь-який текст природною мовою, який анотований тегами, визначеними *TEI*, стає стандартним об'єктом *IT*-середовища, здатним до використання багатьох складних програм. Оскільки тегування особливе для тексту, який кодується, документи *TEI* є унікальними. Хоча певні теги можуть бути однаковими незалежно від джерела, існують інші теги, які є ексклюзивними для різних жанрів, зокрема таких як драма, поезія чи проза. Також розроблено певні шаблони, тобто порожні документи *TEI* з основними наборами тегів. Кожен шаблон може відрізнятися чітким набором тегів, що робить кожен такий документ унікальним. Доступні шаблони можна використовувати або змінювати відповідно до цілей проєкту. Перспективу подальших досліджень вбачаємо в залученні цих технологій у суміжних дисциплінах: мовознавстві, перекладознавстві, літературознавстві.

Список літератури:

1. Демська-Кульчицька О. М. Застосування принципів TEI до кодування текстових корпусних даних. *Проблеми програмування.* 2004. № 2–3 [спец. вип.]. С. 422–430.
2. Муковнін Є. В. Труднощі в роботі систем обробки природної мови та основні методи їх вирішення. *Сучасні філологічні дослідження та навчання іноземної мови в контексті міжкультурної комунікації* (Х). 2017. С. 343–347.
3. Logar N., Kosem I. TERMIS: A corpus-driven approach to compiling an e-dictionary of terminology. *Electronic lexicography in the 21st century: thinking outside the paper*. Proceedings of the eLex 2013 conference. Ljubljana / Tallinn, 2013. P. 164–178.
4. Pinnis M., Gornostay T., Skadiņš R. et al. Online Platform for Extracting, Managing, and Utilising Multilingual Terminology. *Electronic lexicography in the 21st century: thinking outside the paper*. Proceedings of the eLex 2013 conference. Ljubljana / Tallinn, 2013. P. 122–131.
5. Reese R. M. Natural Language Processing with Java. PacktPublishing, 2015. 262 p.
6. Rösiger Ina, Schäfer Jo., George T. et al. Extracting terms and their relations from German texts: NLP tools for the preparation of raw material for specialized e-dictionaries. *Electronic lexicography in the 21st century: linking lexical data in the digital age*. Proceedings of the eLex 2015 conference. Ljubljana / Brighton, 2015. P. 486–503.
7. TEI by Example. URL: https://teibyexample.org/examples/TBED00v00.htm
8. Text Encoding Initiative. URL: https://tei-c.org/
9. What is TEI? URL: https://cdrh.unl.edu/articles/basicguide/TEI

Hrytsiv N. M., Babiak S. A., Sofiianyk N. I. APPLIED ASPECTS OF LINGUOTECHNOLOGIES

The article analyses applied aspects of linguotechnologies. In particular, the core linguotechnological principles of terminological e-dictionaries creation are considered with reference to the foreign experience of computer lexicography, and the review of opportunities for automating the natural language processing is suggested.

From a lexical perspective, the object of scientific interest is due to the increasing need to represent the conceptual and terminological apparatus of applied fields. Since the conceptual system is subjected to arrangement, critical and extensive analysis, clear-cut delineation, terminological dictionaries serve as main means of scientific and technical terminology normalisation. In particular, explanatory translation dictionaries serve as a special kind of the terminological ones.

Therefore, the article identifies basic stages of dictionary generation, such as determining the source base, automating the selection process, pre-processing, term selection, and statistical ranking. In addition, the latest processes of material pre-processing (e. g. tokenisation, lemmatisation, POS-tagging) are singled out, and advanced solutions for automating the terminological data extraction (e. g. an innovative cloud platform TaaS) are considered.

Apart from lexicographic perspectives, applied aspects of linguotechnologies have the potential for processing not individual lexemes, but entire text arrays. It is known that over the past two centuries, mankind has effectively used the automation of numerous tasks using mechanical and electrical technologies. With that in mind, the article also considers the problem of automating the natural language processing, namely the encoding of texts using the TEI (Text Encoding Initiative). In addition, the general description of TEI as an international community of natural language researchers is presented.

Since the problems of natural language processing are far from unambiguous, their study is of great scientific interest. The article also provides a generalised description of the principles of the TEI system. Considerable attention is paid to the review of different sets of markup, namely TEI P3 and TEI P5, their common and distinctive features to describe the structure, appearance and content of the text.

Key words: *applied linguistics, natural language, tokenisation, lemmatisation, markup, terminology database, linguotechnologies, TEI standard, text encoding initiative, tag.*