**УДК 658.012.011.56**

**І.М. Кульчицький, Н.Б. Осідач**
Національний університет "Львівська політехніка",
кафедра прикладної лінгвістики

# КОНЦЕПТУАЛЬНА МОДЕЛЬ БАЗИ ДАНИХ ЗБІРКИ М. НОМИСА "УКРАЇНСЬКІ ПРИКАЗКИ, ПРИСЛІВ'Я І ТАКЕ ІНШЕ"



**В Україні одним із завдань збереження культурної спадщини є переведення в електронну форму, одна з яких – лексикографічна база даних, національних надбань у галузі культури та мистецтва, серед яких чимало лексикографічних праць. До таких надбань належить збірка М.Т. Номиса «Українські приказки, прислів'я і таке інше». Стаття присвячена побудові концептуальної моделі бази даних для цієї збірки.**

**Ключові слова: електронна лексикографія, фразеографія, електронна версія, база даних, фразеологічна одиниця, лінгвокультурологія.**

**One of the tasks of preserving the cultural heritage in Ukraine is the digitalization of the national achievements in the sphere of culture and art which include many lexicographic works. One of their electronic forms is a lexicographic database. The collection "Ukrainian adages, proverbs and the like" by M. Nomys belongs to such national achievements. The article is dedicated to the construction of a conceptual model of the database for this collection.**

**Key words: Electronic lexicography, phraseography, electronic version, database, phraseological unit, linguistic culturology.**

## Вступ

Для сучасної лексикографічної теорії та практики особливої актуальності набули три проблеми [5, 3]:

- Оновлення фундаментальних лексиконів. Швидкий темп сучасного життя спонукає постійні зміни в мові, що, своєю чергою, вимагає оновлення словникової продукції відповідно до вимог часу.
- Необхідність переведення лексикографічної спадщини людства в цифрову форму. Мета – нагромадження загальнокультурного потенціалу та використання її (спадщини) в системах автоматичного опрацювання інформації. Для цього абсолютно недостатньо просто сканувати тексти відповідних словників, хоча й це не така проста задача, як це може здатися з першого погляду. Ефективність роботи з такими, особливо великими, словниками потребує їхньої конверсії у формати лексикографічних баз даних.
- Застосування словників у формуванні лінгвістичних компонент концептографічних систем експлікації знань. Це потребує не лише великих інформаційних масивів, але й спонукає до віднаходження у традиційних словникових текстах прихованих семантичних структур.

Друга проблема особливо важлива для України, адже її вирішення дасть можливість запровадити в широкий науковий обіг максимальну кількість інформації з репертуару виданих лексикографічних праць, особливо тих, які через ті чи інші причини стали бібліографічною рідкістю. І хоча більшість наявної у них інформації втратила свою цінність для широкого загалу, на наше глибоке переконання, вона абсолютно не втратила своєї наукової значущості для науковців різних галузей лінгвістики. У цих словниках законсервоване знання тодішніх вчених, що були належними фахівцями у своїх галузях, та тогочасна українська лексика, частина якої сьогодні забута, але її з успіхом можна повернути в широкий вжиток [2, 322]. Тому Закон України „Про основні засади розвитку інформаційного суспільства в Україні на 2007–2015 роки" до однієї з основних стратегічних цілей розвитку інформаційного суспільства в Україні зараховує збереження культурної спадщини України [6, р. I, п. 1]. З цією метою у законі передбачено [6, р. III, пп. 4, 10]:

- створення в електронній формі архівних, бібліотечних, музейних фондів та інших фондів закладів культури;
- формування відповідних інформаційно-бібліотечних та інформаційно-пошукових систем з історії, культури, народної творчості, сучасного мистецтва України тощо;
- переведення в електронну форму національних надбань у сфері культури та мистецтва.

До таких надбань належить збірка Матвія Терентійовича Номиса (Симонова) «Українські приказки, прислів'я і таке інше»[8], яка має низку особливостей, що роблять її унікальною фольклорною пам'яткою. У ній уперше в українській фразеографії застосовано тематичний принцип та запроваджено чітку паспортизацію одиниць за місцем їхньої фіксації. Перший раз її видано в м. Санкт-Петербурзі (Росія) в 1864 р. з численними купюрами тодішньої цензури. На початку XX ст. у «Записках Наукового товариства ім. Шевченка» (1909, т. 88, кн. 2) опубліковано зразки паремій, вилучені цензурою в першому виданні. За радянських часів у 20-х роках минулого століття збірку перевидав відомий фольклорист й етнограф А. Лобода. У подальшому збірку в радянській Україні не перевидавали до 90-х років минулого століття, коли журнал «Київ» друкує частинами тексти зі збірки Номиса. За кордоном збірку перевидано в 1985 р. За незалежної України збірку в 1993 р. видало київське видавництво «Либідь», розпочавши нею серію «Літературні пам'ятки України», воно ж і перевидало її в 2004 р. в серії «Пам'ятки історичної думки України». Незважаючи на назву „збірка", автори статті зараховують її до фразеологічних словників та розглядають її як лексикографічну працю [2, 321–322].

**Постановка проблеми**

Повернення лексикографічних праць у науковий обіг можливе двома шляхами – перевидання друкованим способом та відтворення в електронній формі. Без сумніву, перевидання повертає такі словники з небуття. Однак з огляду на економічні чинники їх перевидають не так уже й часто та й невеликим накладом, що не сприяє їхньому широкому розповсюдженню. За приблизним підрахунком, до 1948 р. видано близько 214 словників та проектів словників, до складу яких входить українська мова. На сьогодні в серії «Із словникової спадщини» їх перевидано 10 [2, 322].

На нашу думку, продуктивнішим є відтворювати ці лексикографічні і фразеографічні твори в електронній формі. Такий підхід певною мірою дешевший та простіший у тиражуванні. Проте найважливіше – електронна форма урізноманітнює доступ до інформації в цих словниках та підвищує якість її наукового опрацювання.

Вищеподані фактори обумовлюють **актуальність** створення електронної версії збірки М. Номиса „Українські приказки, прислів'я і таке інше".

**Аналіз останніх досліджень та публікацій**

За однією з типологічних схем [4, 141] електронні словники (словники, записані за допомогою електронних пристроїв на електронних носіях інформації) групують за ознаками „засіб використання" та „спосіб доступу". За першою ознакою словники ділять на:

- комп'ютерні – записані та відтворюються на персональному або портативному комп'ютері;
- кишенькові – записані на кишенькових електронних пристроях, наприклад кишенькові перекладачі;
- мобільні – записані в мобільних телефонах.

За другою ознакою виділяють:

- стаціонарні – встановлені на жорсткому диску комп'ютера;
- переносні – записані на компакт-дисках з доступом тільки за наявності в дисководі;
- он-лайн – розміщені на комп'ютері-сервері в мережі й доступні її засобами, наприклад, інтернет.

Переважно ці словники розраховані на широкого споживача, що обумовлює склад їхньої інформаційної бази, критерії пошуку та структуру відповіді на пошукові запити. Окрім того, не завжди можна встановити джерела інформаційної бази таких словникових систем. Вдалими прикладами таких систем можуть слугувати продукти фірми ABBYY Lingvo, розділ «Словари» сайту російського гуманітарного університету (http://www.i-u.ru/biblio/dict.aspx) та проект «Словники України on-line» (http://lcorp.ulif.org.ua/dictua/) Українського мовно-інформаційного фонду НАНУ. Компакт-версія останнього містить засоби автоматизованої лематизації словоформ.

Щодо повернення в науковий обіг лексикографічної спадщини, то в цьому напрямку в Україні активно проводять роботу, на жаль, за невеликими винятками, ентузіасти без жодної державної підтримки. Як правило, лексикографічні та фразеографічні раритети сканують та надають доступ до зісканованої інформації у форматах pdf, djvu, doc (див. сайти http://culture-ua.com/catalog/, http://litopys.org.ua/links/inlex.htm, http://www.mova.info/, http://ukrknyga.at.ua/load/, http://mposhuk.com.ua/item/7891236842258/lang/ru, http://www.ukrlife.org/main/, http://www.madslinger.com/-/bookvault/index.), або забезпечують пошук у зісканованій інформації (прикладом може слугувати сайт http://leksika.com.ua/). На думку авторів, найпліднішe в цьому напрямку працює проект „Російсько-українські словники" [7], де реалізовані обидва способи доступу до інформації. Цікаву збірку паремій подає сайт http://aphorism.org.ua/.

У жодному разі не применшуючи необхідність, важливість та корисність наявних електронних форм (автори залюбки користуються ними у своїй повсякденній роботі), зазначмо, що бракує ще однієї форми – текстів словників кінця XIX – початку XX століття у вигляді інформаційних систем з пошуковим апаратом, який відповідав би сучасним потребам. Це – можливість пошуку слів за сучасною орфографією з одночасним переглядом інформації в орфографії оригіналу, пояснення малозрозумілих сучасному читачеві слів тощо. Ми вважаємо, що таке завдання могли б виконувати кафедри прикладної лінгвістики вищих навчальних закладів України, залучаючи студентів, що опановують цю спеціальність [3].

Саме в такому напрямі на кафедрі прикладної лінгвістики Нац. ун-ту „Львівська політехніка" розпочато створення електронної версії збірки М. Номиса „Українські приказки, прислів'я і таке інше". За основу взято видання збірки у видавництві «Либідь» 1993 року [8]. У попередній статті [2] було проаналізовано структуру збірки, на основі чого було подано її інформаційну модель. Окрім того, було описано технологію перетворення збірки в електронну форму. При цьому автори дотримуються двох основних засад: максимально зберегти автентичність збірки та максимально уможливити сучасним дослідникам наукове опрацювання інформації збірки. На сьогодні текст збірки перетворено у форму, придатну для опрацювання текстовим редактором, електронний текст звірено з оригіналом перший раз (будуть і наступні звіряння) та виправлено виявлені помилки і розроблено систему маркерів для первинної структуризації інформації, які вставлено в текст збірки. **Мета** цієї статті – на основі інформаційної моделі [2, 323–326] запропонувати концептуальну модель лексикографічної бази даних, у яку буде конвертовано текст збірки.

## Основний матеріал

Інформаційну модель збірки відображатимемо в реляційну модель, яка буде використана під час побудови її лексикографічної бази даних. Під час проектування дотримуватимемося рекомендацій, які подані в [1]. Для пояснення структури концептуальної моделі використаємо такі статті збірки:

***58. Бог старий господарь.*** Бр. – *…**має більше, ніж роздасть. – Більше Бог має, як роздав.*** Ил., Кан., К. [8, 41]

***174. Те іде молицьця, а те живицьця (**красти**)**. *Зал.* [8, 45]

***176. Святий та Божий! свічки поів, а поночі сидить.*** *Кон., К., Евх*. *Як не по правді робить, та ще й виправляєцця* [8, 45]

***1063. Як мисль, так мисль – таки буде Перемишль.*** *Ил. Стародавній галицький город над Саном. Славлять, що слова сі сказав цареві простий Русин, як той добірав мення городові* [8, 88]

***2597. Не вважайте***([1])***, люде добрі***, ***що я швець: говоріть зо мною*** ([2]) ***як з простим*** ([3]). *Бр.*– ***Чуєш, говори зо мною, як з простим, а не думай того, що я швець.*** *Не.*

([1]) ***Не вжахайтеся.*** *І. Бр.*; ***Не потурай.*** *Бр., Проск.*; ***Не думай***. *Коз.* ([2]) ***говоріть***. *І. Б р.*; ***говори***. *Бр., Проск., Коз.* ([3]) ***з простим чоловіком.*** *І. Бр., Проск*; ***до простого***. *Бр.* [8, 148]

***2823. Нічим*** ([1]) ***вовкові*** ([2]) ***блювати*** ([3]), ***так*** ([4]) ***ликами*** ([5]) *Пр.*

([1]) ***Немає чим.*** *Ил., Проск.* ([2]) ***вовк***. *Ил., Пр. в Ст. Зб.* ([3]) ***смердіти***. *Проск.*; ***с...ь***. *Л.*, *Новг.*, ***вовка рвать.*** *Гл.; Ил., Л.* ([4]) ***то*** *Ил.*; ***та.*** *Проск.* ([5]) ***ликом***, *id.*; ***завертами*** (***завертнями***). *Новг.* [8, 159]

Інформаційна модель збірки розрізняє три рівні – макро-, медіо- та мікрорівень, подані на рис. 1–3 [2, 324–326].

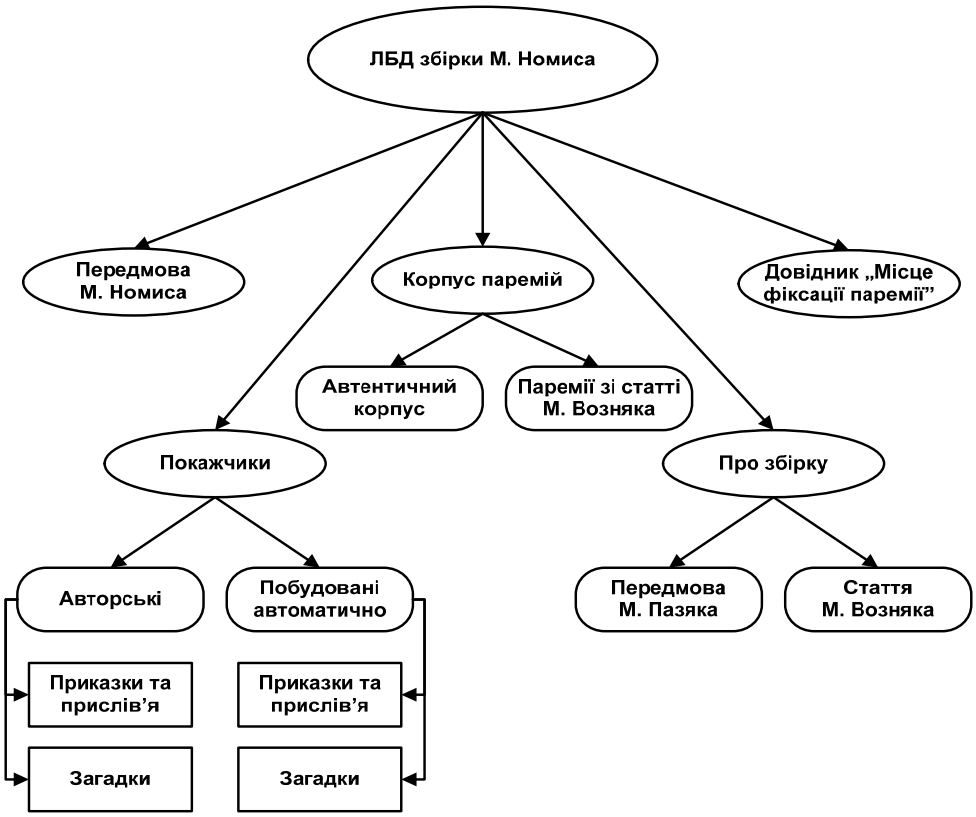


*Рис. 1. Макроструктура збірки М. Номиса*

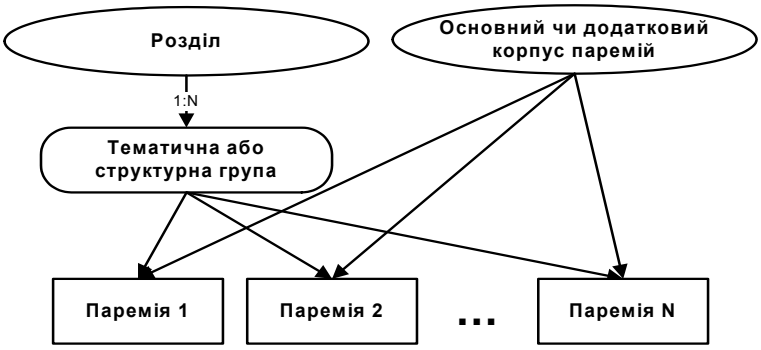


*Рис. 2 Медіоструктура збірки*

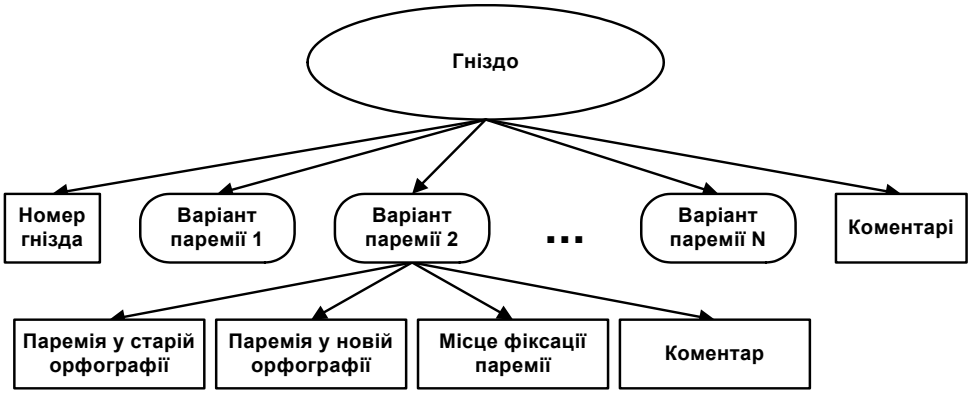


*Рис. 3 Мікроструктура збірки*

Якщо структура відношень макрорівня більш-менш очевидна (надалі буде подана структура відношення „Місце фіксації паремії”), то справа зі структурою відношень медіо- та мікрорівня дещо інша. Проектування розпочнемо, починаючи з найнижчого – мікрорівня. При його відображенні нам необхідно враховувати те, що структура оригіналу повинна бути максимально збережена, та що інформація у статтях буде супроводжуватися сучасними коментарями. Для цього вчинимо так. Весь текст збірки відобразимо у вигляді списку словоформ. Після цього необхідно провести лематизацію та кожну словоформу співвіднести з її канонічною (словниковою) формою. Такий підхід дозволить надалі застосувати зміну старої орфографії на сучасну один раз, та мати гарантію, що для однієї словоформи буде подано єдиний варіант у сучасній орфографії.

За бажання та необхідності лексикографічну базу даних можна буде доповнити тлумаченнями слів, що використані у збірці з метою видання відповідного словника. Отже, отримуємо два відношення «Список слів» та «Список слоформ», структура яких подана на рис. 4. У випадку наших прикладів вміст відношень буде таким (з метою економії місця вміст відношень подано лише для паремії за номером 2597 (див. вище):

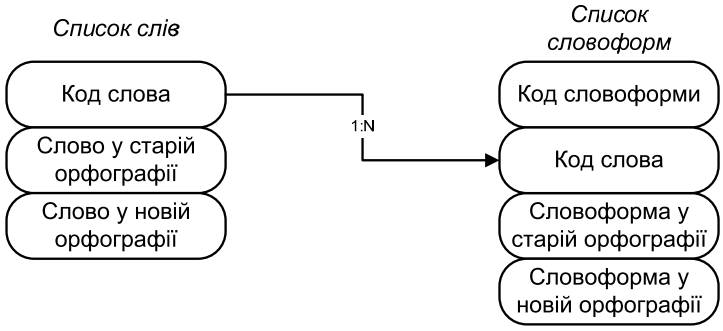


*Рис. 4 Структура відношень “Список слів” та “Список словоформ”*

*Таблиця 1*

**Відношення “Список слів”**

| Код слова | Слово в старій орфографії | Слово в новій орфографії |
|---|---|---|
| 1 | а | а |
| 7 | вважати | вважати |
| 8 | вжахатися | вжахатися |
| 12 | говорити | говорити |
| 15 | до | до |
| 17 | добрий | добрий |
| 18 | думати | думати |
| 20 | з | з |
| 23 | зо | зо |
| 29 | людина | людина |
| 34 | не | не |
| 42 | потурати | потурати |
| 44 | простий | простий |
| 63 | той | той |
| 66 | чоловік | чоловік |
| 67 | чути | чути |
| 68 | швець | швець |
| 70 | що | що |
| 71 | я | я |
| 72 | як | як |

*Таблиця 2*

**Відношення “Список словоформ”**

| Код словоформи | Код слова | Словоформа в старій орфографії | Словоформа в новій орфографії |
|---|---|---|---|
| 1 | 1 | а | а |
| 7 | 7 | вважайте | вважайте |
| 8 | 8 | вжахайтеся | вжахайтеся |
| 14 | 12 | говори | говори |
| 15 | 12 | говоріть | говоріть |
| 19 | 15 | до | до |
| 21 | 17 | добрі | добрі |
| 22 | 18 | думай | думай |
| 24 | 20 | з | з |
| 27 | 23 | зо | зо |
| 33 | 29 | люде | люди |
| 37 | 71 | мною | мною |
| 40 | 34 | не | не |
| 48 | 42 | потурай | потурай |
| 51 | 44 | простим | простим |
| 52 | 44 | простого | простого |
| 74 | 63 | того | того |
| 78 | 66 | чоловіком | чоловіком |
| 79 | 67 | чуєш | чуєш |
| 80 | 68 | швець | швець |
| 82 | 70 | що | що |
| 83 | 71 | я | я |
| 84 | 72 | як | як |

Наступним кроком утворюємо два відношення для подання блоків слів, які містяться у збірці. Під блоком розумітимемо частину паремії, яка або незмінна, або має варіанти, зафіксовані у різних місцевостях, або коментар М. Номиса. Наприклад, для паремій 1063 та 2597 це буде (номери блоків відповідають вмісту описаного нижче відношення «Список блоків» для цих паремій):

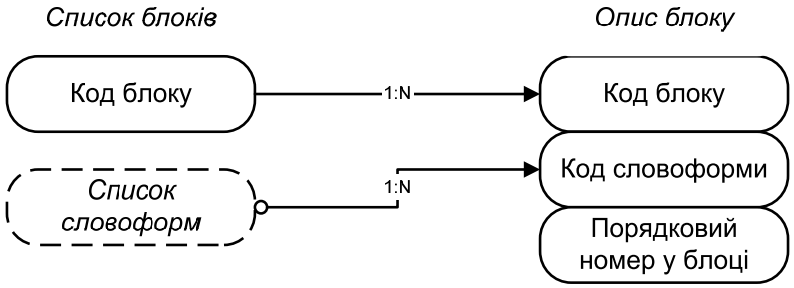


*Рис. 5 Структура відношень для подання блоків тексту*

1. ***Як мисль, так мисль – таки буде Перемишль.***
2. ***Стародавній галицький город над Саном. Славлять, що слова сі сказав цареві простий Русин, як той добірав мення городові***
3. ***Не вважайте***
4. ***Люде добрі, що я швець***
5. ***Говоріть зо мною***
6. ***Як***
7. ***З простим***
8. ***Чуєш, говори зо мною, як з простим, а не думай того, що я швець.***

***9. Не вжахайтеся.***

***10. Не потурай.***

***11. Не думай.***

***12. Говоріть.***

***13. Говори***

***14. З простим чоловіком***

***15. До простого.***

Структуру цих відношень подано на рис. 5.

На цьому рисунку в овалі, який зображено пунктирною лінією, показано відношення „Список словоформ", структура якого подана на рис. 4. Вміст відношення „Опис блока" для паремій з номером 1063 та 2597 буде таким (для економії місця його подано в дві колонки):

*Таблиця 3*

**Вміст відношення "Опис блока"**

| Код блока | Код словоформи | Порядковий номер у блоці | Код блока | Код словоформи | Порядковий номер у блоці |
|---|---|---|---|---|---|
| 3 | 40 | 1 | 8 | 1 | 8 |
| 3 | 7 | 2 | 8 | 40 | 9 |
| 4 | 33 | 1 | 8 | 22 | 10 |
| 4 | 21 | 2 | 8 | 74 | 11 |
| 4 | 82 | 3 | 8 | 82 | 12 |
| 4 | 83 | 4 | 8 | 83 | 13 |
| 4 | 80 | 5 | 8 | 80 | 14 |
| 5 | 15 | 1 | 9 | 40 | 1 |
| 5 | 27 | 2 | 9 | 8 | 2 |
| 5 | 37 | 3 | 10 | 40 | 1 |
| 6 | 84 | 1 | 10 | 48 | 2 |
| 7 | 24 | 1 | 11 | 40 | 1 |
| 7 | 51 | 2 | 11 | 22 | 2 |
| 8 | 79 | 1 | 12 | 15 | 1 |
| 8 | 14 | 2 | 13 | 14 | 1 |
| 8 | 27 | 3 | 14 | 24 | 1 |
| 8 | 37 | 4 | 14 | 51 | 2 |
| 8 | 84 | 5 | 14 | 78 | 3 |
| 8 | 24 | 6 | 15 | 19 | 1 |
| 8 | 51 | 7 | 15 | 52 | 1 |

Наступним кроком створюємо відношення, яке описує гніздо паремії. Структура його подана на рис. 7. Дамо необхідні пояснення.

Номер паремії – збігається з номером паремійного гнізда у збірці.

Підномер паремії – використовується, якщо варіанти у гнізді подано блоком.

Код блоку – ключовий атрибут з відношення „Список блоків".

Порядковий номер блока в паремії – описує порядок блоків у паремії.

Номер варіативності – для блоків, які описують варіант паремії, збігається з номером у збірці, для основної паремії – „0".

Підномер варіативності – використовують, якщо під одним номером варіанту вказано декілька варіацій з різних місцевостей

Тип блока – „0" паремія, „1" – коментар

Код місця фіксації паремії – ключовий атрибут відношення „Місце фіксації паремії", структуру якого подано нижче:


*Рис. 6. Структура відношення "Місце фіксації паремії"*


*Рис. 7. Структура відношення "Гніздо паремії"*

Для паремії 2597 вміст відношення "Гніздо паремії" буде таким:

*Таблиця 4*

**Вміст відношення "Гніздо паремії"**

| Номер паремії | Підномер паремії | Код блока | Порядковий номер блока в паремії | Номер варіативності | Підномер варіативності | Тип блока | Код місця фіксації паремії |
|---|---|---|---|---|---|---|---|
| 2597 | 1 | 3 | 1 | 1 | 0 | 0 | Бр |
| 2597 | 1 | 4 | 2 | 0 | 0 | 0 | Бр |
| 2597 | 1 | 5 | 3 | 2 | 0 | 0 | Бр |
| 2597 | 1 | 6 | 4 | 0 | 0 | 0 | Бр |
| 2597 | 1 | 7 | 5 | 3 | 0 | 0 | Бр |
| 2597 | 2 | 8 | 1 | 0 | 0 | 0 | Не |
| 2597 | 3 | 9 | 1 | 1 | 1 | 0 | І. Бр |
| 2597 | 4 | 10 | 1 | 1 | 2 | 0 | Бр |
| 2597 | 4 | 10 | 1 | 1 | 2 | 0 | Проск |
| 2597 | 5 | 11 | 1 | 1 | 3 | 0 | Коз |
| 2597 | 6 | 12 | 1 | 2 | 1 | 0 | І. Бр |
| 2597 | 7 | 13 | 1 | 2 | 2 | 0 | Бр |
| 2597 | 7 | 13 | 1 | 2 | 2 | 0 | Проск |
| 2597 | 7 | 13 | 1 | 2 | 2 | 0 | Коз |
| 2597 | 8 | 14 | 1 | 3 | 1 | 0 | І. Бр |
| 2597 | 8 | 14 | 1 | 3 | 1 | 0 | Проск |
| 2597 | 9 | 15 | 1 | 3 | 1 | 0 | Бр |

Наступним кроком побудуємо відношення, які описують розподіл паремій за категоріями, запропонованими авторами. Структура цього відношення така (рис. 8):

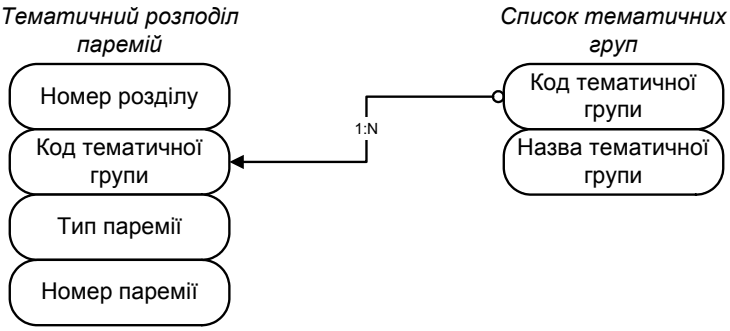


*Рис. 8. Структура відношень, які описують тематичний розподіл паремій*

У відношенні „Тематичний розподіл паремій” атрибут „тип паремії” вказано на те, чи паремія ввійшла у збірку, чи перебуває у списку, який опублікував М. Возняк.

**Висновки**

Запропоновано концептуальну модель лексикографічної бази даних збірки М. Номиса «Українські приказки, прислів’я і таке інше». Елементи цієї моделі можуть бути використані під час створенні електронних версій видань схожого типу.

*1. Дейт К. Дж. Введение в системы баз данных / К. Дж. Дейт.– 8-е изд.– М.: Вильямс, 2006.– 1328 с. 2. Кульчицький І.М., Осідач Н. Б. Лексикографічна база даних збірки М. Номиса “Українські приказки, прислів’я і таке інше” / І.М. Кульчицький, Н. Б. Осідач // Вісник: Інформаційні системи та мережі: Зб. наук. праць. – Львів: Нац. ун-т "Львівська політехніка", 2010, № 699. – С. 321–331. 3. Кульчицький І.М. Бази даних як засіб повернення у науковий обіг української лексикографічної спадщини / І.М. Кульчицький, І.О. Ліхнякевич, Е.А. Калниня // Українська термінологія і сучасність: Зб. наук. праць / Відп. ред. Л. О. Симоненко. – К.: Інститут української мови НАН України, 2009. – Вип. 8. – С. 273–275. 4. Кульчицький І. М. Деякі аспекти типології словників / І. Кульчицький, А. Костенко, Н. Осідач // Прикладна лінгвістика та лінгвістичні технології: MegaLing-2008: зб. наук. пр. / НАН України, Укр. мовно-інформ. фонд [та ін.]; редкол.: Ю.Д. Апресян [та ін.]. – К.: Довіра, 2009. – С. 135–145. 5. Лінгвістичні та технологічні основи тлумачної лексикографії / В.А. Широков, В.М. Білоноженко, О.В. Бугаков та ін. – К.: Довіра, 2010. – 295 с. 6. Про основні засади розвитку інформаційного суспільства в Україні на 2007-2015 роки: закон України від 9 січня 2007 року № 537-V // Відомості Верховної Ради України. – 2007. – № 12. – С. 102. 7. Російсько-українські словники [Електронний ресурс]. – Режим доступу: http://www.r2u.org.ua/ 8. Українські приказки, прислів’я і таке інше. Уклав М. Номис / Упоряд., приміт. та вступна ст. М. М. Пазяка. – К.: Либідь, 1993. – 768 с. («Літературні пам’ятки України»). – К. : Либідь, 2004. – 352 с. («Пам’ятки історичної думки України»).*