УДК 81'33

МОДЕЛЬ АНОТАЦІЇ ТЕКСТОВОГО КОРПУСУ ЯК ЗАСІБ ДОСЛІДЖЕННЯ ХУДОЖНЬОЇ КАРТИНИ СВІТУ

Гладкова Ганна Павлівна,
асп.
Київський національний університет імені Тараса Шевченка

Статтю присвячено розробці трирівневої моделі анотації лексичних одиниць художнього тексту, яка дозволяє проводити комплексне вивчення мовного явища у єдності його семантики, синтактики та прагматики, що у свою чергу дає змогу вийти на рівень аналізу картини світу художнього твору.
Ключові слова: *корпус, анотація, XML, запит.*

Сучасні мовознавчі дослідження зосереджені на вивченні реального функціонування мовних одиниць, що зумовлює високу актуальність методів корпусної лінгвістики – одного з найперспективніших напрямів комунікативно-функціональної парадигми. Корпусні методи з успіхом застосовуються у комп'ютерній лексикографії, методиці викладання іноземних мов, перекладознавстві, і останнім часом набувають популярності і у когнітивних дослідженнях мови [Gries 2006]. ***Метою*** даної статті є запропонувати модель анотації художнього матеріалу для проведення аналізу мовних явищ та описати її практичну реалізацію на мові метаопису XML. ***Об'єктом*** дослідження є параметри лексичних одиниць (англійських абстрактних іменників із суфіксом *–ness*), що описують їх функціонування у художньому тексті, а ***предметом*** – модель анотації, що їх містить.

Корпус можна визначити як "сукупність мовних даних, представлених у електронному вигляді" [Гвишиани 2008, 5] або як "колекцію текстів... які зберігаються у електронній базі даних" [Baker 2006, 48]. У такому розумінні корпусом можна назвати будь-яку сукупність електронних текстів, до яких є спільний інтерфейс доступу. Корпус не може бути універсальним, тобто таким, у якому достатньо повно репрезентовані всі регістри мови у всі періоди її існування. Проте більшість лінгвістичних корпусів сконструйовано з тим, щоб якомога повніше репрезентувати певний тип дискурсу у заданий період, у такому обсязі, який є необхідним і достатнім для всебічного вивчення поставленої проблеми. Вибір матеріалу та його подальша обробка також зумовлюються метою конкретної роботи.

Корпуси можна поділити на дві великі групи: корпуси, що містять "чистий" текст, та корпуси, що містять анотований текст, тобто спеціально структурований та підготовлений таким чином, щоб забезпечити досліднику можливість опрацьовувати потрібний саме йому матеріал. Корпуси, що містять "чисті" тексти, можуть бути придатні для швидкої перевірки певної гіпотези у дослідженнях обмеженої лексичної групи або таких мовних явищ, які можуть бути описані за допомогою регулярних виразів – словотвірних типів, алітерації, деяких граматичних явищ тощо (хоча пошук за регулярними виразами не гарантує відсутність помилок).

Проте лінгвістичний аналіз зазвичай передбачає введення класифікаційних схем та вивчення мовних явищ як категорій, тобто їх опис за певними параметрами. Сучасні технології дозволяють додавати необхідну інформацію безпосередньо до лінгвістичного матеріалу таким чином, у який пошук необхідної групи прикладів мо-

же бути здійснений автоматично. Процес додання "службової" інформації до тексту називається анотуванням, а його результат – анотацією або розміткою. Анотовані корпуси, у свою чергу, за форматом зберігання текстових даних можна поділити на декілька груп. Широко використовується так званий "вертикальний формат" (у якому слово з тексту та інформація, яка його стосується, розміщуються на одному рядку через табуляцію). Окремий випадок становлять метамови, які дозволяють представити текстову інформацію у структурованому вигляді. Так, активно використовується мова опису SGML (Standard Generalized Mark-up Language), розроблена у 1986 році, але на даному етапі її витісняє її різновид під назвою XML (Extended Markup Language), запропонований організацією W3 Consortium у 1998 році [Bray 2006]. На даний момент XML є максимально гнучкою метамовою, яка широко використовується у різноманітних базах даних, але серед мовознавців, окрім спеціалістів з прикладної лінгвістики, вона ще мало відомий. Тому коротко зупинимося на її структурних елементах.

Файл XML є звичайним текстовим файлом у кодировці UTF-8, що може бути підготовлений у будь-якому текстовому редакторі (існують і спеціалізовані програми, здатні перевірити правильність структури такого документу). Файл XML містить текст, структурні частини якого відокремлюються за допомогою спеціальних елементів – тегів, взятих у кутових дужки. Початок та кінець елементу позначають початковим та кінцевим тегом (наприклад, слова звичайно розділяють тегом <w>: <w>cat</w>). Тегі утворюють ієрархічну структуру, тобто один тег може містити один чи більше інших тегів. Тегі можуть містити атрибути, у яких зручно додавати інформацію про параметри мовної одиниці (наприклад <w semantic_category="animal">cat</w>). Файл у форматі XML обов'язково містить стандартний заголовок <?xml version="1.0" encoding="UTF-8"?> та кореневий елемент – тег, у який включені всі інші теги.

Оскільки назви, зміст та кількість тегів та атрибутів ніяк не обмежені, мова XML дає змогу додати до тексту будь-яку інформацію залежно від цілей конкретного дослідження. У даній статті запропоновано модель анотації, розроблену для когнітивно-дискурсивного аналізу словотвірного типу іменників із суфіксом *–ness* на матеріалі англомовної художньої прози.

Когнітивно-дискурсивна парадигма мовознавства передбачає всебічне вивчення мовного явища у його функціонуванні у дискурсі і його лінгвокультурну інтерпретацію. У даному дослідженні аналіз словотвірного типу проводилося на рівні цілого художнього тексту у тріаді семантика-синтактика-прагматика. Тексти діахронічного корпусу романів XVIII-XX сторіч обсягом близько 3,5 мільйонів слів було відібрано за тематичним принципом та за критерієм впливовості (influentialness) [Kennedy 1998, 63], тобто перевага надавалася творам, які мають велику культурну значимість. Джерелами текстів корпусу слугувала електронна бібліотека університету Аделаїди (http://www.ebooks.adelaide.edu.au).

Обробка матеріалу проводилася у декілька етапів. Слова із суфіксом *–ness* досить просто відшукати за допомогою регулярних виразів, але спочатку потрібно виключити слова з омонімічними закінченнями (governess, witness та інші). Потім

тексти було автоматично поділено на параграфи та речення за допомогою програми "Sptoolkit" Скота Піао [Piao 2008] (відповідно теги <s> та <p>). Цей етап не є обов'язковим, але анотація речень та параграфів може бути корисною для пошуку таких контекстів, у яких створюється стилістичний ефект за рахунок декількох уживань досліджуваного явища.

Цілі дослідження зумовлюють анотування іменників із суфіксом *–ness* на трьох рівнях – семантики, синтактики та прагматики. З цього списку лише анотацію синтаксичного рівня можна здійснити автоматично. Оскільки інформація про прагматичний контекст слова додається вручну, доцільно спочатку обробити цей рівень, а потім скористатися програмами семантичного та морфологічного аналізу, які може прийняти вже анотований вхідний текст.

Продемонструємо цей процес на прикладі речення з "Гордості та упередження" Дж.Остін. Для анотації іменників на *–ness* вводимо тег **<ness>**.

```
<p><s>"Mr. Darcy is all politeness," said Elizabeth, smiling.</s></p>
```

Прагматичний рівень. У інформацію прагматичного рівня входять насамперед два пункти: хто говорить та про що говорить. Вони позначаються атрибутами **speaker** та **qualified_item** відповідно.

```
<s>"Mr. Darcy is all <ness sem_group="polite" speaker="Elizabeth"
qualified_item="Darcy">politeness</ness>," said Elizabeth, smiling.</s>
```

Семантичний рівень. Всі слова у тексті проходять процедуру токенізації (тобто всі слова відокремлюються одне від одного тегами <w> та нумеруються за допомогою атрибутів **id** – ідентифікаторів окремих лексичних одиниць (у прикладі, наведеному нижче, ці атрибути опущені для економії місця)). У даному дослідженні семантичний аналіз здійснюється на базі системи USAS, розробленої у Ланкастерському університеті [Archer 2004]. Через велику кількість помилок та особливості концептуалізації за допомогою іменників із суфіксом *–ness* результати цього етапу вимагають верифікації. Атрибут **sem** містить інформацію про семантичний клас слова (наприклад, Z1f є умовним позначенням жіночих імен).

```
<p><s> <w sem="PUNC">"</w><w sem="Z1m[i324.2.1">Mr.</w>
<w sem="Z1m[i324.2.2">Darcy</w><w sem="A3+">is</w><w sem="N5.1+">all</w>
<ness sem_group="polite" speaker="Elizabeth" qualified_item="Darcy">
<w sem="S1.2.4+"> politeness</w></ness> <w sem="PUNC">,</w>
<w sem="PUNC">"</w><w sem="Q2.1">said</w> <w sem="Z1f">Elizabeth</w>
<w sem="PUNC">,</w><w sem="E4.1+">smiling</w><w sem="PUNC">.</w></s></p>
```

Синтаксичний рівень. Для визначення морфологічного класу слів (так звані pos-tags) використовується система морфологічної розмітки CLAWS [Garside 1987], розроблена у Ланкастерському університеті. Атрибут **pos** містить інформацію про його морфологічний клас (наприклад, NP1 є умовним позначенням власних іменників, а VVD – дієслова з повним лексичним значенням у минулому часі). Саме цю

програму було використано у підготовці Британського національного корпусу; її точність на публіцистичних текстах сягає приблизно 93 %. Після проведення цього етапу вищенаведене речення виглядатиме таким чином:

```
<p><s><w pos="YQUO" sem="PUNC">"</w><w pos="NNB" sem="Z1m[i324.2.1">
Mr.</w><w pos="NP1" sem="Z1m[i324.2.2">Darcy</w><w pos="VBZ" sem="A3+">
is</w><w pos="DB" sem="N5.1+">all</w><ness sem_group="polite"
speaker="Elizabeth" qualified_item="Darcy">
<w pos="NN2" sem="S1.2.4+">politeness</w></ness>
<w pos="," sem="PUNC">,</w> <w pos="YQUO" sem="PUNC">"</w>
<w pos="VVD" sem="Q2.1">said</w> <w pos="NP1" sem="Z1f">Elizabeth</w>
<w pos="," sem= "PUNC">,</w> <w pos="VVG" sem="E4.1+">smiling</w>
<w pos="." sem="PUNC">.</w></s></p>
```

Запропонована модель анотації представляє всі типи інформації, внесені у атрибути тега <ness>, як категорії, і тому дає змогу створювати запити як на окремих рівнях семантики, синтактики та прагматики, так і на їх перехресті. Наприклад:

⇨ *семантика:* які семантичні класи іменників із суфіксом *–ness* присутні у корпусі або окремому тексті;

⇨ *синтаксис:* у складі яких дистрибутивних моделей уживаються іменники із суфіксом *–ness* і чи є вони ідіостилістично зумовленими;

⇨ *прагматика:* які іменники уживаються для концептуалізації відношень між певними персонажами;

⇨ *семантика-синтаксис:* чи спостерігається тяжіння між окремими семантичними класами іменників із суфіксом *–ness* та певними типами дистрибутивних моделей;

⇨ *семантика-прагматика:* які персонажі частіше застосовують слова певної семантики;

⇨ *синтаксис-прагматика:* які типи модальності тяжіють до яких дистрибутивних моделей, які синтаксичні конструкції уживають ті чи інші персонажі, яка середня довжина речень у окремих персонажів.

Після того, як анотацію корпусу закінчено, можливо створювати запити до нього. Стандартним засобом запитів до баз даних XML є мова запитів XQuery, розроблена W3 Consortium [Boag 2007]. Проте ця мова є складною для гуманітарних спеціалістів і не підтримує істотні особливості текстових корпусів (зокрема на даний момент не підтримуються порядок елементів у базі даних, відстань між ними, а також регулярні вирази). З XML-корпусами можливо працювати за допомогою програми Xaira, розробленої для нового видання Британського національного корпусу [Aston, Burnard 2006]. До її недоліків слід віднести необхідність індексації корпусу, певну складність та нестабільність у роботі. Для паралельних корпусів можливо також користуватися програмою Alinea [Duchet J.-L. 2005]. У рамках даного дослідження було взято участь у проекті створення спеціалізованої програми для роботи з XML-корпусами невеликого обсягу [Gladkova, Drozd 2009], яка не потребує індексації та реалізує просту мову запитів на основі зразків. Поточна (консольна) версія програми опублікована за веб-адресою: http://sourceforge.net/projects/xcorp.

Отже, у даній статті запропонована трирівнева модель анотації лексичних одиниць художнього тексту, яка дозволяє проводити комплексний аналіз мовного явища у єдності його семантики, синтактики та прагматики. Дана модель є гнучкою та може бути адаптована до досліджень різних груп лексичних одиниць та модифікована у залежності від завдань конкретної роботи. Її можливо застосовувати також у корпусах текстів інших регістрів, компаративних дослідженнях різних типів дискурсу тощо.

Статья посвящена разработке трехуровневой модели аннотации лексических единиц художественного текста, которая позволяет провести комплексное изучение исследуемого языкового явления в единстве его семантики, синтактики и прагматики и, таким образом, выйти на уровень анализа художественной картины мира.

Ключевые слова: *корпус, аннотация, XML, запрос.*

The paper describes a three-level annotation scheme for the lexical corpus designed for use in studies of language registers. The presented scheme incorporates semantics, syntax and pragmatics of lexical items, thus enabling the researcher to study the role they play in language world model arised from literary texts.

Keywords: *corpus, annotation, XML, query.*

Література:

1. *Гвишиани Н.Б.* Практикум по корпусной лингвистике: Учебное пособие по английскому языку / Н.Б. Гвишиани. – М., Высшая школа, 2008. – 191 с.

2. *Archer D. et al.* The UCREL semantic analysis system [Electronic resource] / Archer D., Piao S., Rayson P., McEnery T. // Proc. of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks in association with 4th International Conference on Language Resources and Evaluation (LREC 2004), 25th May 2004, Lisbon, Portugal. – Paris: ELRA, 2004. – pp. 7-12. – http://www.comp.lancs.ac.uk/computing/users/paul/publications/usas_lrec04ws.pdf.

3. *Aston G., Burnard L.* Introducing XAIRA: an XML-aware concordance program [Electronic resource] / Guy Aston, Lou Burnard. – Presentation at workshop held at TALC 2006. – http://www.oucs.ox.ac.uk/rts/xaira/Talks/xaira-wkshop.odp.

4. *Baker P.* A Glossary of Corpus Linguistics / Paul Baker, Andrew Hardie, Tony McEnery. – Edinburgh: Edinburgh University Press, 2006. – 187 p.

5. *Boag S. et al.* XQuery 1.0: An XML Query Language [Electronic resource] / Scott Boag, Don Chamberlin, Mary F. Fernández. – 2007. – http://www.w3.org/TR/xquery.

6. *Bray T. et al.* Extensible Markup Language (XML) 1.1 (Second Edition) [Electronic resource] / T. Bray, J. Paoli, C. M. Sperberg-McQueen. – http://www.w3.org/TR/2006/REC-xml11-20060816.

7. *Duchet J.-L.* Alinea: a language independant tool for bi-text processing [Electronic resource] / Jean-Louis Duchet, Oliever Kraif // JRC EU-Enlargement Workshop: Exploiting parallel corpora in up to 20 languages. JRC-Ispra, Italy, 26-27.09.2005. – http://langtech.jrc.it/0509_EU-Enlargement-Workshop.html.

8. *Garside R.* The CLAWS Word-tagging System [Electronic resource] / Roger Garside // The Computational Analysis of English: A Corpus-based Approach. – London: Longman, 1987. – http://ucrel.lancs.ac.uk/papers/ClawsWordTaggingSystemRG87.pdf.

9. *Gladkova G.P., Drozd A.A.* Towards easier querying of XML-based linguistic corpora / G.P. Gladkova, A.A. Drozd // Таврический Вестник Информатики и Математики. – 2009. – № 2. – С. 71-77.

10. *Gries S.* Corpora in cognitive linguistics: Corpus-based approaches to syntax and lexis / S. Gries (ed). – Berlin: Walter de Gruyter, 2006. – 352 p. – (Trends in linguistics. Studies and monographs; 172).

11. *Kennedy G. D.* An Introduction to Corpus Linguistics / Graeme Kennedy. – London: Longman, 1998.

12. *Piao S.* A Highly Accurate Sentence and Paragraph Breaker [Electronic resource] / Scott Piao, 2008. – http://text0.mib.man.ac.uk:8080/scottpiao/sent_detector.