УДК 81-13: 81'322.5

Соломія Бук (м. Львів)

СТРУКТУРНЕ АНОТУВАННЯ У КОРПУСІ ТЕКСТІВ (НА ПРИКЛАДІ ПРОЗИ ІВАНА ФРАНКА)

Стаття презентує деякі результати, отримані в процесі укладання корпусу текстів І. Франка. Зокрема, уточнено поняття структурне анотування корпусу текстів. Виявлено важливі елементи структурного анотування корпусу прозових текстів І. Франка, а саме: авторське визначення жанру твору, примітки автора про переробку, присвята, епіграф, цитата, авторський підпис твору (дата і/або місце написання, правдиве ім'я чи псевдонім), авторські та редакторські покликання. Теги на їх позначення можна вважати необхідними для кожного корпусу текстів окремого автора.

Ключові слова: *корпус текстів, структура тексту, структурне анотування корпусу текстів, тег.*

У сучасній лінгвістиці створюють і використовують корпуси текстів (КТ) для вирішення найрізноманітніших завдань: від дослідження морфології та лексичного значення слів [32] до лінгвістичної експертизи тексту [1] і метафори [31]. Проте розвиток КТ окремих авторів перебуває на початковій стадії. Твори конкретних письменників в комп'ютерному варіанті існують швидше як електронні бібліотеки (Бібліотека української літератури [2]; The online books page («Онлайнова книжкова сторінка») [38]; Проект Gutenberg [39]) або конкорданси (Конкорданція поетичних творів Тараса Шевченка) [16]; конкорданси до творів українських поетів другої половини XX століття (І.Драча, М. Вінграновського, Л. Костенко, І. Калинця та ін.) [26]; он-лайн конкорданс роману І. Франка "Перехресні стежки" [4]; Конкорданс публіцистики Ф. Достоєвського [22]; Конкорданс усіх творів В. Шекспіра [30]; Ранговий конкорданс роману "Уліс" Дж. Джойса [44]). Така форма представлення робить легшим доступ до текстів, уможливлює створення частотних списків слово-

форм, дає можливість знаходити конкретні слова, словосполуки чи фразеологізми. Але вона не може замінити КТ як "машиночитане, стандартно організоване зібрання репрезентативних для певної мови, діалекту або іншої підмножин(и) мов(и) писемних або усних текстів, призначених для лінгвістичного аналізу й опису, відібраних і впорядкованих згідно з експліцитними екстра- та інтралінгвальними критеріями" [10: 53]. "Довільне зібрання електронних текстів не можна вважати корпусом у термінах корпусного мовознавства, якщо таке зібрання не має певних ознак, а саме: репрезентативності, збалансованості, автентичності, комп'ютерної підтримки, документованості та стандартності" [10: 75]. Як правило, у КТ розрізняють представлення позамовної інформації (*зовнішнє анотування*: дані про автора (стать, вік, освіта тощо) та твір (час написання, виходу друком, редакцію тощо)) та внутрішньомовної інформації (*внутрішнє анотування*: морфологічні, лексико-семантичні, синтаксичні дані тощо) Див., наприклад, [12: 73].

Таким чином, КТ конкретного автора дає можливість глибинно вивчати мову письменника, його стиль на орфографічному, морфологічному, лексичному, логіко-семантичному, фразеологічному, синтаксичному, прагматичному, дискурсивному та інших рівнях. Це — необхідне джерело для сучасних академічних видань і перевидань повного зібрання творів письменника, для укладання словника мови письменника, для порівняння різних (у т. ч. прижиттєвих) варіантів того самого твору і т. д.

КТ конкретного автора має деякі особливості у порівнянні із загальномовним КТ на усіх рівнях анотації, включно зі структурним. Так, наприклад, для загальномовного корпусу, створеного для лексикографічних потреб, необов'язкова вказівка на усі структурні елементи твору (епіграф, присвята тощо), тоді як для письменницького корпусу вони є необхідними.

У науковій літературі віддавна виявляли зацікавлення до закономірностей побудови тексту. У передмові до [11] одним із факторів, що зумовив такий інтерес, названо інформаційний вибух, яким характеризується науково-технічна революція. Він "вимагає різкого підвищення ефективності обміну інформацією, що неможливо до використання ЕОМ, а для цього необхідно дати ЕОМ параматери побудови тексту" [11: 3].

Стаття презентує деякі результати, одержані у процесі роботи над укладанням корпусу текстів Івана Франка (КТФ). У другому розділі уточнено дефініцію терміносполуки *структурне анотування* для корпусу текстів письменника. Розділ 3 містить опис структурного анотування КТ слов'янських мов. У Розділі 4 коротко викладено концептуальні засади анотування КТФ, а у п'ятому подано практичне застосування структурного анотування на КТФ, визначено його особливості, запропоновано теґи для оригінальних структурних елементів тексту І. Франка.

Уточнення дефініції терміносполуки *структурне анотування*

Анотованість — це та характеристика, яка відрізняє КТ від електронного збору текстів. Оскільки в Україні корпусна лінгвістика — порівняно новий напрям, її термінологічний апарат перебуває на стадії становлення. Тому паралельно послуговуються такими термінами, як розмітка, анотування, маркування, індексування на позначення процесу та результату присвоєння текстовим одиницям маркерів, в яких закодовано певну лінгвістичну інформацію.

Терміносполука "структурна анотація" у спеціальній літературі має неоднакові дефініції. Найширше її розуміння знаходимо у Ч. Меєра: "Структурне анотування передбачає описову інформацію про текст. Наприклад, ..."заголовок файлу".., ...повний бібліографічний опис писемного тексту або етнографічна інформація про учасників (напр., їх вік і стать) в усному діалозі. ...Можна врахувати додаткове структурне анотування для позначення ...меж параграфу в писемному тексті або накладання сегментів мовлення в усному тексті" [36: 81].

Як видно із цитати, дослідник використовує цю терміносполуку на позначення і структури тексту, і зовнішньої стосовно нього інформації (його бібліографічний опис, дані про мовців тощо). У вужчому значенні "структурну анотацію" описують Ґ. Астон і Л. Бернард: "...корисно вказувати межі глав, розділів, абзаців, речень, і т. д., а також особливу роль заголовків, переліків, приміток, посилань, супровідних підписів, покликів та ін." [27: 24].

Міжнародний стандарт кодування текстової інформації TEI (*англ.* Text Encoding Initiative — проект кодування текстів), який "репрезентує ті ознаки тексту, які потребують експліцитної ідентифікації з метою сприяти текстовому опрацюванню за допомогою комп'ютерних програм" [43], під елементами універсальної структури тексту пропонує розуміти <head> (заголовок), (частина, розділ), <p> (абзац), <s> (речення), <epigraph> (епіграф), <dateline> (дата), <note> (примітка), <said> (пряма мова), <dedication> (присвята), <l> (рядок, у вірші), <abbr> (скорочення), <num> (число) та ін. Аналогічно структурну анотацію інтерпретують у проектах створення корпусів українських текстів: Український національний лінгвістичний корпус [17: 207 — 208, 275], Національний корпус української мови [10: 152 — 168]. Так само цю терміносполуку розуміють автори корпусу російської публіцистики кінця XIX століття [9] та Національного корпусу російської мови [19].

Отже, структурою тексту вважаємо такі його елементи як назва, розділ, підрозділ, рубрика, присвята, епіграф, поклик, цитата, вживання алфавітів інших писемних систем, цифр тощо. Структурне анотування КТ — це виділення структурних елементів тексту за допомогою певної мови маркування; сукупність маркерів-вказівок на елементи зовнішньої будови тексту.

Структурне анотування у практиці деяких корпусів текстів

Важливо зазначити, що на сьогодні проблема створення повного КТ конкретного автора перебуває на початковій стадії розвитку. Можна знайти інформацію про існування КТ В. Шекспіра [37], Ф. Достоєвського [22], В. Шевчука [18] та ін., проте самі ці корпуси у вільному користуванні недоступні. Тому звернемося до загальних КТ національних мов. Важливою ознакою КТ є стандартність мови анотування. На сьогодні найбільш поширеним стандартом кодування інформації в електронних текстах став TEI [43]. Тому багато укладачів КТ взяли за основу систему маркерів саме цього міжнародного міждисциплінарного стандарту, яка дає можливість робити деякі модифікації набору теґів у кожному конкретному випадку залежно від мети створення КТ. Так, наприклад, Korpus języka polskiego PWN (Корпус польської мови Польського наукового видавництва при Польській Академії наук), створений для вдосконалення джерельної бази словників та полегшення лексикографічної роботи, не має спеціальних маркерів для присвяти, епіграфа, підпису, оскільки їх вважали зайвими стосовно завдання цього КТ [33].

Český národní korpus (Чеський національний корпус) [29], Slovenský národný korpus (Словацький національний корпус) [42], Национальный корпус русского языка (Національний корпус російської мови) [19], Korpus Języka Polskiego IPI PAN (Корпус польської мови Інституту основ інформатики Польської Академії наук) [33] також не вирізняють цих характеристик, трактуючи їх як необов'язкові для репрезентації національних мов. Детальніший опис КТ польської мови див. також [41]. На подібних засадах побудовані системи розмітки українських текстів, що незалежно створюють у двох осередках: Інституті української мови НАНУ [10] та Українському мовно-інформаційному фонді НАНУ [17].

У схемі маркування, яку використовують у Проекті кількісного аналізу текстів (Quantitative Text Analysis, QuanTA) Університету м. Ґрац (Австрія), деякі споріднені теґи об'єднано: [iad] (Interne Addenda) для епіграфів, покликань, кінцевих приміток, звертань (у листах) та ін. Окремий теґ [att] (Autortexttyp) позначає авторську класифікацію тексту [40]. Отже, поданий огляд КТ підтверджує факт, що вибір тегів і форми їх подання залежить від тих дослідницьких завдань, для розв'язання яких створюють корпус.

Загальні засади анотування Корпусу текстів Івана Франка (КТФ)

Корпус текстів окремого автора можна створити, виокремлюючи тексти письменника із національного КТ, якщо їх туди залучено. Проте загальний КТ із дуже високою ймовірністю буде містити не всі твори письменника. Отже, укладений таким чином КТ письменника не буде репрезентативним, а це неприпустимо, оскільки репрезентативність — одна з обов'язкових вимог до будь-якого КТ.

КТФ заплановано як репрезентативний машиночитаний анотований статичний дослідницький. Основні концептуальні параметри його створення викладено у [3]. Проект створення КТФ передбачає, що корпус буде охоплювати усі твори І. Франка (а це орієнтовно 7 млн. слововживань) з усіма особливостями фонетичного, морфологічного, словотвірного, лексико-семантичного, фразеологічного, синтаксичного рівнів. КТФ репрезентує підсистему західного варіанту української мови кінця XIX — початку XX ст. І, що видається особливо важливим, не тільки у художньому, а й у публіцистичному, науковому (галузь літературознавства, етнографії, культурології, економіки, філософії, соціології), епістолярному стилях. Величезний та різнорідний обсяг спадщини письменника зумовив поділ КТФ на декілька підкорпусів: велика та мала проза, драматургія, літературно-критичні праці, публіцистика, наукові розвідки, епістолярій, поезія, перекладні твори.

У 1960-х рр. у Львівському університеті був задум створити словник мови І. Франка, теоретичне підґрунтя якого сформулював проф. І. Ковалик [13, 14]. Було створено картотеку, яка нині зберігається в Івано-Франківському педінституті [20: 82]. Одним із результатів роботи колективу авторів під керівництвом професора став Словопокажчик поетичних творів Івана Франка [15]. Отже, поезія І. Франка вже частково була об'єктом лексикографічного опису з елементами статистики. Тому виконання нашої роботи розпочато з підкорпусу прози, як одного з найбільших за обсягом, який добре репрезентує багатство та різноманітність мови письменника (тільки в одному романі "Перехресні стежки" зафіксовано 9 978 різних слів (!), тоді як Словник мови Т. Шевченка у двох томах містить 10 116 слів [6]).

КТФ — дослідницький корпус, зорієнтований на широкий клас лінгвістичних завдань, при цьому, що більше параметрів буде включено в його анотування, то більший його дослідницький потенціал. Сподіваємося, він стане зручним інструментом в опрацюванні ідіостилю письменника, оскільки на всьому обсязі творів І. Франка дасть можливість здійснювати автоматичний пошук будь-якої мовної одиниці (наприклад, лексеми, конструкції, граматичної форми) у будь-якій формі у всіх контекстах; здійснювати автоматичний пошук твору за будь-яким параметром (наприклад, твори певного жанру, написані певного року (періоду), певною мовою, підписані певним псевдонімом, з епіграфом чи без нього тощо). Оскільки твори у КТФ датовані, вдасться фіксувати час першого та останнього використання певної лексеми в текстах, прослідкувати хронологію та динаміку розвитку мови автора. Це значною мірою уможливить реконструювання західного варіанта української літературної мови кінця XIX — поч. XX ст.; елементів територіальних та соціальних діалектів цього ж періоду, адже своїх персонажів автор наділяв мовленням того середовища, яке вони представляли. За допомогою КТФ стане можливим отримати, окрім якісної, кількісну характеристику мови письменника (на основі

корпусу можна автоматично укладати частотні словники і конкорданси до будь-якого конкретного чи групи заданих творів) тощо. Саме корпус дасть можливість зробити максимально ефективною роботу над укладанням словника мови письменника, а також над порівняльним вивченням різних редакцій, версій і т. п. конкретних творів.

У КТФ за типом інформації (позатекстова і власне текстова) виділено два основних рівні анотування: зовнішній та внутрішній. На зовнішньому подано зовнішню стосовно тексту інформацію, метадані: бібліографічний опис творів, вік та освіта автора тощо. На внутрішньому рівні подано внутрішню стосовно тексту інформацію: структурну (яка стала об'єктом опису цієї статті), морфологічну, синтаксичну, семантичну. Окремого опрацювання вимагає анафоричне та прагматичне анотування. Подібні принципи опису текстів є й у інших КТ, зокрема у Корпусі російської мови XIX ст. [9], див. також [21]. Зараз уже виконано морфологічне анотування деяких творів, результатом чого стали, зокрема он-лайн конкорданс [4] та частотний словник роману "Перехресні стежки" [5], а також виявлено основні статистичні параметри цього тексту [28].

Відповідно до усталеної в мовознавстві традиції, в основу КТФ лягли останні прижиттєві видання творів І. Франка, а також найповніше на сьогодні Зібрання творів у 50-ти томах [24] і видання творів, що до нього не ввійшли [25]. Самі твори письменника мають глибоку багатошарову будову, яка у перевиданнях додатково ускладнюється редакторськими примітками та покликаннями, перекладами фрагментів тексту іншими мовами та поясненнями. Тому у процесі створення КТФ виявлено особливі структурні елементи тексту письменника, не описані у попередньо згаданій теоретичній літературі, а саме, у найбільш поширеному стандарті кодування інформації в електронних текстах TEI, а також у найбільших КТ слов'янських мов, згаданих вище. До таких, наприклад, належать авторське визначення жанру твору, авторський підпис тексту (дата і/або місце написання твору, справжнє ім'я або псевдонім), авторські примітки про переробку тощо.

Особливості структурного анотування корпусу текстів Івана Франка

Загальними засадами до набору теґів для будь-якого КТ є багаторазове використання, сумісність з іншими КТ та сумісність із загальноприйнятими науковими теоріями. Тому за основу анотування КТФ було взято принципи міжнародного стандарту TEI, які до того ж відповідають обов'язковій вимозі уніфікованості КТ.

Загальна схема структурного анотування КТФ:

Заголовок позначатимемо теґом <head>, частини, розділи — теґами або <div0>, <div1>, <div2> і т. д., якщо твір має кількарівневий поділ. Для відзначення ***підзаголовків*** пропонуємо ввести теґ <subhead>. Головною одиницею основного тексту є абзац, для якого використовуватимемо стандартний теґ <p>. Межі речень виділятимемо теґами <s>...</s>.

Зважаючи на мету створення КТФ і з урахуванням того, що передмова чи післямова становлять з художнім текстом єдине ціле, пропонуємо не виокремлювати ці структурні елементи поза межі основного тексту <body>, як це зроблено в TEI, де передмова належить до зони <front>, а післямова — до <back>. Для відзначення ***передмови*** використовуватимемо тег <foreword>, а для ***післямови*** — <afterword>. Нижче наведено анотування початкових сторінок повісті "Захар Беркут", де використано також елементи, зміст яких описано далі.

```
<head>Захар Беркут</head>
<subhead>
 Образ громадського життя
  Карпатської Русі в XIII віці
</subhead>
<foreword n="Передмова">Передмова
<p>
<s>Повість історична -- се не історія.</s>
...
<s>Дійові особи зрештою видумані, місцевість списана по можливості вірно.</s>
</p>
<dateline>
 <place>Львів</place>, <date>дня 1 грудня 1882</date>.
</dateline>
<byline><author>Ів. Франко</author></byline>
...
<remade>
 <dateline>
  <place>Криворівня</place>, <date>1 серпня 1902</date>.
 </dateline>
 <byline><author>І. Ф.</author></byline>
</remade>
</foreword>

<epigraph>
 <lang="ru">
 <lg>
  <l>Дела давно минувших дней,</l>
  <l>Преданья старины глубокой...</l>
 </lg>
 <byline><author>А. С. Пушкин</author></byline>
 </lang>
</epigraph>

<div0 n="I">I
<p>
<s>Сумно і непривітно тепер в нашій...
```

У попередньому прикладі групу віршованих рядків, відповідно до TEI, відзначено теґом <lg> (line group), а окремі віршовані рядки — <l>. Наприклад, ще [24 XIV, 153]:

```
<lg>
 <l>Кажуть люде, що суд буде,</l>
 <l>А суду не буде!</l>
 <l>Най на того суд упаде,</l>
 <l>Хто судити буде!</l>
</lg>
```

Цитування. На противагу багатьом словникам письменника, наприклад [7; 8], які не включають у свій опис цитати з творів інших авторів або іншомовні вставки до тексту, було вирішено, що КТФ повинен охоплювати всі текстові елементи творів автора, оскільки комп'ютерні програми зможуть їх визначати автоматично. Це відкриє нові перспективи для сучасних лінгвістичних досліджень, наприклад, які твори, яких авторів, яких національностей і якими мовами цитував І. Франко, поезію чи прозу, фольклор чи авторські роботи. Цей комплекс проблем може стати об'єктом окремого дослідження. Цитування позначатимемо теґом <cit>.

Авторське визначення жанру власного твору (як правило, зазначене відразу після заголовка) є дуже важливим, оскільки подає авторський план і оцінку твору. Із цього погляду для укладання КТФ є щонайменше дві проблеми:

1) Авторське визначення жанру не збігається із сучасним. Наприклад, Іван Франко називає "Перехресні стежки" *повістю* [24 XX: 173], хоча сучасні спеціалісти-літературознавці визначають його як *роман*.

2) Автор створив оригінальний жанр, наприклад, *Мавка* (Літня казочка) [24 XV: 91 — 95]. Подібний структурний елемент виникає, коли назва містить деякі біографічні факти: *У столярні* (Із моїх споминів) [24 XXI: 171 — 188]. Інколи на цьому місці Іван Франко вказував на головного героя або тему твору, наприклад, *Полуйка* (Оповідання старого ріпника) [24 XX: 7 — 25], *Терен у нозі*. Оповідання з гуцульського життя [24 XXI: 375 — 390].

Для вказівки на авторське визначення жанру пропонуємо теґ <agd> (Author's genre definition). Цей елемент разом з деякими іншими, згаданими вище, проілюструємо на кількох прикладах:

```
<head>Перехресні стежки</head>
<subhead><agd>Повість</agd></subhead>
<div0 n="I">I
<p>
<s><said speaker="Стальський">-- А, пан меценас!</s>
...
<head>Мавка</head>
<subhead><agd>(Літня казочка)</agd></subhead>
```

…

<head>Моя стріча з Олексою</head>
<subhead>(<agd>оповідання</agd> <pseudo>Мирона Сторожа</pseudo>)</subhead>

…

<head>У столярні</head>
<subhead>(Із моїх споминів)</subhead>

…

Додаткова інформація (зазвичай, час, дата подій і под.) може міститися на початку розділу (тоді вона належить до теґа <opener>) або наприкінці (<closer>). Наприклад [24 XV: 199],

<head>Із записок недужого</head>
…

<opener>
 <time>9 година</time>
</opener>
<p><s>Грав трохи в варцаби…

Авторський підпис твору може містити зазначення місця та часу написання, ім'я автора (повне: *І. Франко, Іван Франко, Dr. Iwan Franko*; ініціали: *І. Ф.*; псевдонім: *Мирон сторож* та ін.). У деяких творах авторство вказане на самому початку, наприклад, *Ріпка* // Стара казка, по-новому розповів І[ван] Ф[ранко] [24 XVII: 306 — 307]:

<head>Ріпка</head>
<subhead>
 Стара <agd>казка</agd>, по-новому розповів
 <author> І[ван] Ф[ранко]</author>
</subhead>

Інколи у структуру тексту включено авторські примітки про переробку. Для них ми ввели теґ <remade>. Скажімо, оповідання *Цигани* завершується таким підписом [24 XVI: 166]:

Нагуєвичі, в липні 1882,
перероблено 1887.

Його вмістимо в таку структуру:

<closer>
 <dateline>
  <place>Нагуєвичі</place>,
  <date>в липні 1882</date>
 </dateline>,
 <dateline>
  <remade>перероблено <date>1887</date></remade>.
 </dateline>
</closer>

Примітки (самого автора, редакторів видання) позначатимемо відповідно <note resp=”auth”> і <note resp=”editor”>. Наприклад, в оповіданні “Яндруси” [24 XVII: 212 — 213] маємо кілька авторських пояснень, включно із заголовком:

<head>Яндруси<note resp=”auth”>В жаргоні львівських вуличників -- хлопці</note></head>

...

<s>Войцехова рано була десь там на хрестинах, вернула аж о одинадцятій і то вже під доброю датою<note resp=”auth”>Підохочена, напідпитку</note>.</s>

...

Редакторські примітки зазвичай супроводжують іншомовні вставки у тексті, наприклад, в оповіданні “Свинська конституція” [24 XX: 7–13]:

<head>Свинська конституція</head>
<dedication>
Присвячую пам’яті
Антона Грицуняка
</dedication>

...

<s>Грицуняк -- се дуже цікава поява, один із немногих живих іще<note resp=”auth”>Він умер 29 марта 1900 р.</note> недобитків того племені оповідачів ...

...

... і там так твердо навчилися “s&lslash;oma-siano”<note resp=”editor”>Солома-сіно (польськ.).-- Ред.</note>, ...

...

(Останньої примітки немає в публікації 1900 року у збірці “Сім казок” [23]. &lslash; позначає польську літеру ł.)

Нестандартне розділення підрозділів знаками * [24 XIV: 117 тощо] або — [24 XIV: 278 тощо] і подібні позначатимемо так:

*

Якщо певні структурні елементи відбиті від інших порожніми рядками (за винятком віршованих вставок), то, залежно від логічного навантаження, використовуватимемо <div n=” ”> для елементів, близьких за змістом до підрозділів, або <emptyline> в інших випадках (напр., відділення тексту листів [24 XVII: 446, 450 тощо]).

Із викладеного стає зрозуміло, що КТ конкретного письменника на структурному рівні має свої особливості у порівнянні із загальномовним КТ. Відповідно до вимоги репрезентативності структурне анотування повинно охоплювати всі без винятку елементи будови тексту. Зокрема, авторське визначення жанру твору, підпис твору, вказівку на час, місце його створення чи переробки.

Запропоновані у статті доповнення та деякі модифікації міжнародного стандарту TEI дають можливість більш повно та адекватно відобразити глибинну організацію й архітектоніку тексту конкретного письменника, відрізнити авторське коригування та пояснення тексту від редакторського тощо. Ці факти набувають особливого значення у світлі підготовчих робіт до видання повного зібрання творів І. Франка у 100 томах (задум Інституту літератури ім. Т.Г.Шевченка НАН України). Електронне автоматичне опрацювання текстів Франкової спадщини, зокрема описане анотування, значно полегшить мовознавчі, літературознавчі, герменевтичні, текстологічні та історичні дослідження франкознавців. Розглянуті структурні елементи тексту можна вважати необхідними як для КТФ, так і для кожного КТ окремого автора загалом.

УМОВНІ СКОРОЧЕННЯ

КТ — корпус текстів
КТФ — корпус текстів Івана Франка
TEI — Text Encoding Initiative

1. *Баранов А.Н.* Лингвистическая экспертиза текста: Учеб. пособие. — М.: Флинта: Наука, 2007. — 592 с.
2. Бібліотека української літератури. — [Цит. 25 жовтня 2008]. — Доступно з: <http://www.ukrlib.com.ua>
3. *Бук С.* Корпус текстів Івана Франка: спроба визначення основних параметрів // Горизонти прикладної лінгвістики. Доп. міжнарод. наук. конференції 20–27 вересня 2006, Україна, Крим, Партеніт / Ред. В. А. Широков, С. С. Дікарєва. Мовно-інформаційний фонд України. Таврійський нац. ун-т ім. В. І. Вернадського. — Сімферополь: В-во "ДиАйПи", 2006. — С. 115—116.
4. *Бук С., Ровенчак А.* Он-лайн конкорданс роману Івана Франка "Перехресні стежки". — [Цит. 30 вересня 2008]. — Доступно з: <http://www.ktf.franko.lviv.ua/~andrij/science/ Franko/concordance.html>.
5. *Бук С., Ровенчак А.* Частотний словник повісті І. Франка "Перехресні стежки" // Стежками Франкового тексту (комунікативні, стилістичні та лексичні виміру роману "Перехресні стежки") / Ф.С. Бацевич (наук. ред.), С.Н. Бук, Л.М. Процак, А.А. Ровенчак, Л.Ю. Сваричевська, І. Л. Ціхоцький. — Львів: Видав. центр ЛНУ імені Івана Франка, 2007. — С. 138—369.
6. *Ващенко В.С.* (ред.) Словник мови Шевченка. — К.: Наук. думка, 1964. — Т. 1. — 484 с.; Т. 2. — 566 с.
7. *Ващенко В.С., Петрова П.О.* Шевченкова лексика. Словопокажчик до поезій Т. Г. Шевченка. — К.: Видав. Київського держ. ун-ту ім. Т. Шевченка, 1961. — 106 с.
8. *Виноградов В.* (ред.): Словарь языка Пушкина: В 4 тт. / Отв. ред. акад. АН СРСР В.В. Виноградов. — М.: Азбуковник, 2000. — с.
9. *Волков С. С., Захаров В. П.* Параметры описания текстов для корпуса русского языка XIX века // Междунар. конференция "Корпусная лингвистика 2004": Тез. докл. 12 — 14 октября 2004 г., Санкт-Петербург.
10. *Демська-Кульчицька О.* Основи національного корпусу української мови.— К.: Інститут української мови НАН України, 2005. — 219 с.
11. Закономерности структурной организации научно-реферативного текста / За ред. В. С. Перебийніс. — К.: Наук. думка, 1982. — 322 с.
12. *Карпіловська Є.А.* Вступ до комп'ютерної лінгвістики. — Донецьк: Юго-Восток, 2003. — 184 с.
13. *Ковалик І.І.* Наукові філологічні основи укладання і побудови Словника мови художніх творів Івана Франка // Українське літературознавство. Іван Франко. Статті і матеріали. — Львів, 1972. — Вип. 17. — С. 3—10.

14. *Ковалик І.І.* Принципи укладання Словника мови творів Івана Франка // Українське літературознавство. Іван Франко. Статті і матеріали. — Львів, 1968. — Вип. 5. — С. 174 — 183.
15. *Ковалик І.І., Ощипко І.Й., Л.М.* Полюга (уклад.) Лексика поетичних творів Івана Франка. Методичні вказівки з розвитку ялексики. — Львів: ЛНУ, 1990. — 264 с.
16. Конкорданція поетичних творів Тараса Шевченка / Ред. і упоряд.: Олег Ільницький, Юрій Гавриш. У 4 тт. — Торонто, 2001.
17. Корпусна лінгвістика / В. А. Широков, О. В. Бугаков, Т. О. Грязнухіна та ін. — К.: Довіра, 2005. — 471 с.
18. *Монахова Т.В.* Мова Валерія Шевчука: ключові концепти, корпус, тезаурус: Автореф. дис. ...канд. філол. наук. — К.: КНУ, 2008. — 18 с.
19. Национальный корпус русского языка. — [Цит. 30 сентября 2008]. — Доступный с: <http://www.ruscorpora.ru>.
20. *Ощипко І.Й.* Про укладання словника мови поетичних творів Івана Франка // Іван Франко і світова культура. Матеріали Міжнар. симпозіуму ЮНЕСКО (Львів, 11 — 15 вересня 1986). — Кн. 1. — К.: Наук. думка, 1990. — С. 81 — 83.
21. *Рахилина Е.В.* О лексических базах данных // Вопросы языкознания. — 1994. — № 4. — С. 107 — 113.
22. Словарь-конкорданс публицистики Ф.М. Достоевского. — [Цит. 30 сентября 2008]. — Доступный с: <http://dostoevskii.karelia.ru>
23. *Франко І.* Свинська конституция // Сім казок. — Львів: Накл. Українсько-Руської видавничої спілки, 1900. — С. 39–49.
24. *Франко І. Я.* Зібрання творів. У 50 тт. — К.: Наук. думка, 1979.
25. *Франко І. Я.* Мозаїка: Із творів, що не ввійшли до Зібрання творів у 50 тт. / Упоряд. З. Т. Франко, М. Г. Василенко. — Львів: Каменяр, 2002. — 432 с.
26. Частотний словник сучасної поетичної української мови / наук. кер. *Н. П. Дарчук.* – [Цит. 02 жовтня 2003]. — Доступно з <http://www.philolog.univ.kiev.ua/wins/ chast/chast.htm>.
27. *Aston G., Burnard L.* The BNC Handbook. Exploring the British National Corpus with SARA. — Cambridge: Edinburgh University Press, 1998. — 250 p.
28. *Buk S., Rovenchak A.* Statistical Parameters of Ivan Franko's Novel *Perekhresni stežky* (*The Cross-Paths*) // Quantitative Linguistics. — V. 62: Exact Methods in the Study of Language and Text. — Berlin; New York, 2006. — P. 39 — 48.
29. Český národní korpus. — [Цит. 30 вересня 2008]. — Доступно з: <http://ucnk.ff.cuni.cz>
30. Concordance of Shakespeare's complete works. — [Cited 02 October, 2008]. — Available from: <http://www.opensourceshakespeare.com/concordance>.
31. Corpus-Based Approaches to Metaphor and Metonymy / Ed. by *Anatol Stefanowitsch, Stephan Th. Gries.* — Trends in Linguistics. Studies and Monographs 171 / Editors *Walter Bisang* (main editor for this volume), *Hans Henrich Hock, Werner Winter.* — Mouton de Gruyter: Berlin; New York, 2006. — 319 p.
32. *Jones S.* Antonymy: A corpus-based perspective. — London; N. Y.: Routledge, 2002. — XVI, 193 p.
33. Korpus Języka Polskiego IPI PAN.— [Cyt. 2008, 30 września]. — Dostępny z: <http://korpus.pl/index.php?page=welcome>
34. Korpus Języka Polskiego Wydawnictwa Naukowego PWN. — [Cyt. 2008, 30 września]. — Dostępny z: <http://korpus.pwn.pl>.
35. *McEnery T.* Corpus Linguistics // The Oxford Handbook of Computational Linguistics. — Oxford University Press, 2003. — P. 448 — 464.
36. *Meyer C.F.* English Corpus Linguistics: An Introduction. — Cambridge: Cambridge Univetsity Press, 2002. — 168 s.
37. Moby Shakespeare: The complete unabridged works of Shakespeare. — Cited 30 September 2008. — Available from: <http://www.clres.com/corp.html>.
38. Online books page. — Cited 20 October 2008. — Available from: <http://onlinebooks.library.upenn.edu>;
39. Project Gutenberg. — Cited 20 October 2008. — Available from: <http://www.gutenberg.org>
40. QuanTA: Quantitative text analysis. — Cited 02 October 2008. — Available from: <http://www-gewi.uni-graz.at/quanta/>

41. *Rudolf M.* Metody automatycznej analizy korpusu tekstów polskich. — Warszawa: Uniwersytet Warszawski, 2004. — 152s.
42. Slovenský národný korpus. — [Цит. 30 вересня 2008]. — Доступно з <http://korpus.juls.savba.sk/index.sk.html>.
43. TEI: Text Encoding Initiative. P5: Guidelines for Electronic Text Encoding and Interchange. — 2007. — Cited 11 September 2008. — Available from: <http://www.tei-c.org/Guidelines/P5/>
44. Ulysses by James Joyce. A Ranked Concordance. — [Cited 15 October, 2008]. — Available from: <http://www.doc.ic.ac.uk/~rac101/concord/texts/ulysses/ulysses_ranked.html>.

Solomija Buk

STRUCTURE ANNOTATION IN TEXT CORPUS
(ON THE MATERIAL OF IVAN FRANKO'S PROSE)

The paper presents some results which were received while the text corpus of Ivan Franko is being created. The notion of the structure annotation of text corpus is specified. The important text elements for particular structure annotation of an author text corpus are fined out. Such elements as author's definition of genre, author's notes about remaking, dedications, epigraphs, citations, author's signature of work (date and/or place of the work creation, original name or pseudonym), author's and editor's footnotes are described on the material of Ivan Franko's prose.

Keywords: text corpus, text structure, structure annotation of text corpus, tag.

Відомі постаті про мову

Чи завжди південноруська мова була в зневазі – притчею во язицех? Чи завжди розмовляли нею тільки ті, котрі інакше говорити не навчились, писали тільки ті, котрі інакше писати не вміли? Ні!

Ні! У заповітах наших батьків, у ковчезі нашої слави, в історичних свідченнях нашої законорожденності, словом, у наших літописах сильно і часто пробивається справжня південноруська мова ніби з-під кори мови мертвої, священної, церковнослов'янської. Та й як цьому не бути? Розсадник нашої православної віри і нашої народної освіти – південноруський Київ – був і місцем проживання наших перших літописців, і як житейська мова, мова південноруська мимоволі вривалася в їхні дієписання, хоча вони й не підносили їх до Бога під час богослужіння.

Зауважимо, що найдавніший список Нестора, список Лаврентіївський, найбільше міг би нам подати південноруських слів і виразів; але візьмемо навіть Псковський літопис (див. видання Погодіна, Москва, 1837), котрий є скороченням Новгородського, а цей, у свою чергу, є скороченням Київського, і відзначимо (за вказівкою самого п. Погодіна) дещо незрозуміле в ньому без знання південноруської мови...

Та не тільки в Південній Русі, в Червонорусії і Малорусії, панувала південноруська мова і була надбанням як найпростішого, так і найвищого за освітою класу людей; південноруська мова панувала ще і в руських землях, що належали до польського королівства, нею розмовляли й при дворі великих князів литовських і в найбільш знатних домах; нею послуговувались при дипломатичних стосунках, нею писалися важливі державні акти (докази цього можна частково знайти в "Литовській метриці", що зберігається у Варшавському коронному архіві, деякі документи з якого вміщено у "Сборнике" Муха нова). Та й не дивно, якщо поміркуємо, що народонаселення Литви все руське, крім 1/10 частини власне литовського народонаселення.

Є підстави думати, що й сама польська мова кращі риси й багатство своє позичила з мови південноруської, тому що в той час, коли польські письменники думали й писали по-латині, південноруська мова була вже досить розвиненою.

Метлинський А.Л., 1839