УДК 811.581
DOI https://doi.org/110.32838/2710-4656/2021.5-2/06

Ситник І. В.
Київський національний університет імені Тараса Шевченка

ОСОБЛИВОСТІ СТВОРЕННЯ КОРПУСУ ТЕКСТІВ ПІДРУЧНИКІВ КИТАЙСЬКОЇ МОВИ ПОЧАТКОВОГО РІВНЯ

Стаття присвячена дослідженню основних принципів формування корпусу текстів підручників китайської мови початкового рівня. Описано функції корпусу текстів у лінгвістиці. Авторкою відзначені особливості китайської мови, які необхідно враховувати під час створення корпусу текстів китайської мови. Запропоновано основні етапи створення корпусу текстів підручників китайської мови. Систематизовано та охарактеризовано основні принципи створення корпусу текстів: репрезентативність, автентичність, відібраність, збалансованість, машиночитаність. З'ясовано, що процес укладання корпусу пов'язаний також із проблемою виділення тексту, який слід включити в його склад. Проаналізовано змістові принципи відбору текстів власне навчальних підручників: принцип цілісності, принцип відповідності проблемній області, принцип структурної спрямованості. Зважаючи на той факт, що китайська мова належить до кореневого типу мов, встановлено, що у процесі анотування корпусу текстів китайською мовою за основу береться морфологічний аналіз. У статті зосереджено увагу на процедурі поділу тексту на слова та словосполучення, що називається сегментацією, яка створює основу для розмітки. Розглянуто основний набір тегів частиномовної розмітки в корпусах китайської мови. Описаний кожний етап створення власного корпусу текстів підручників китайської мови початкового рівня. На основі охарактеризованих принципів формування корпусу текстів та змістових принципів відбору текстів, а також принципів розмітки та сегментації тексту китайською мовою здійснено конвертування розмічених текстів у структуру спеціалізованої лінгвістичної інформаційно-пошукової системи «Онлайн-корпус китайської мови – автоматична морфологічна розмітка та сегментація тексту». У статті наведені результати дослідження принципів формування корпусу текстів підручників китайської мови початкового рівня та на їх основі сформовано власний машиночитаний, збалансований, репрезентативний, розмічений (анотований) корпус текстів підручників китайської мови початкового рівня для їх кількісного вивчення і якісного пояснення отриманих даних.

Ключові слова: *корпус текстів, китайська мова, початковий рівень, навчальний текст, принципи, підручник.*

Постановка проблеми. Нині в Україні не створено корпусу текстів підручників китайської мови початкового рівня, придатного для вирішення цілої низки лінгводидактичних завдань, а також застосовування як основи для створення вітчизняних підручників, посібників і навчальних матеріалів із китайської мови. Для вирішення цієї проблеми необхідно визначити принципи створення корпусу текстів підручників китайської мови початкового рівня і згідно зі встановленими принципами сформувати відповідний корпус.

Аналіз останніх досліджень і публікацій. Проблема створення корпусу текстів була і залишається предметом інтересу багатьох дослідників, серед яких можна виділити таких, як: Н. Дарчук, В. Широков, В. Жуковська, А. Зубов, А. Карамнов, А. Баранов, В. Захаров, Ґо Шулунь (郭曙纶), Бай Ваньїн (白婉莹), Хуан Цзін (黄婧), Пен Мінь (彭敏), Ян Даньдань (杨丹丹) та інші. Навчальні тексти підручників, зокрема їх структура, неодноразово розглядалася в теорії і практиці навчання іноземних мов, до прикладу, в роботах В. Бейлінсона, Е. Гельфмана, М. Холодної, Цао Маньвень (曹漫雯), Чжан Ніннін (张宁宁), Ван Цюаньлін (王泉玲), Чжен Луяо (郑路遥), Цзян Цзиїн (江子莹), Лі Сяолін (李小玲), Лю Їн (刘颖), Фен Цай (丰彩), Лю Хишань (刘诗涵), Чжоу Сяоє (周晓晔) Ґен Чжи (耿直) та ін. Однак, попри розмаїття наукових праць, котрі вивчають різні аспекти текстового наповнення підручників китайської мови, досі недостатня увага приділена принципам створення корпусу текстів цих підручників, і насамперед – початкового рівня.

Постановка завдання. Мета нашої розвідки полягає в дослідженні основних принципів формування корпусу текстів підручників китайської мови початкового рівня та створенні власного корпусу.

Виклад основного матеріалу. Дослідженням, розроблення, створенням та використанням текстових корпусів, а також лінгвістичним аналізом на їх основі займається корпусна лінгвістика [15, с. 85].

Сьогодні за допомогою корпусу текстів вирішується ціла низка лінгвістичних завдань: реалізація лексико-граматичного аналізу тексту, виявлення термінів і термінологічних словосполучень, складання різноманітних словників тощо. Використання корпусу текстів значно підвищує не лише ефективність і швидкість обробки мовних даних, але й їх достовірність [2, с. 72].

У межах завдань нашої роботи будемо послуговуватися визначенням поняття «корпус текстів», запропонованим проф. Н. Дарчук: «Корпус текстів – це інформаційний лінгвістичний ресурс, в якому користувач може автоматично створити вибірку мовних одиниць та контекстів їхнього вживання на матеріалі текстів певної мови, представлених в електронному варіанті» [7, с. 26].

У типології мовних корпусів виділяють педагогічні та дослідницькі корпуси.

Педагогічні корпуси – це сукупність текстів, що використовуються в освітньому процесі і можуть включати в себе академічні підручники і навчальні посібники, письмові записи комунікації, яка відбувається під час уроків, а також будь-які інші письмові тексти або записи усної мови, які виникають в освітньому середовищі [2, с. 73].

Дослідницькі корпуси широко використовуються у практиці лінгвістичних досліджень і призначені переважно для вивчення різних аспектів функціонування мовної системи і зорієнтований на широкий клас лінгвістичних завдань [4, с. 112].

Таким чином, корпус текстів підручника – окремий вид педагогічно-дослідницького корпусу, оскільки він включає в себе тексти навчальних підручників і розроблений з метою вирішення лінгвістичних завдань, тому процедура його створення буде визначатися факторами, що впливають на створення більш широких типів корпусів даних, до яких він належить [10, с. 59].

Важливою особливістю корпусу текстів, на думку В. Захарова, є те, що він створюється не просто як безліч випадковим чином об'єднаних текстів тієї чи іншої мови, а згідно з певними принципами [1, с. 65]. Під час створення корпусу текстів підручників китайської мови початкового рівня ми дотримувалися принципів, які умовно можна розділити на *принципи формування корпусу* та *змістові принципи відбору текстів*.

До основних принципів формування корпусу текстів належать: репрезентативність, автентичність, відібраність, збалансованість, машиночитаність [9, с. 53].

Репрезентативність полягає в здатності корпусу відображати всі властивості предметної галузі, тобто рівень реалізації мовної системи, що підлягають лінгвістичному описові. У корпусній лінгвістиці можна говорити про кількісну і якісну моделі репрезентативності. Кількісна модель репрезентативності будується на основі частоти явищ у проблемній області. Це означає, що створюваний корпус повинен відображати всі властивості проблемної області (наприклад, мова підручника китайської мови) у певній пропорції. Якісна модель репрезентативності ґрунтується на встановленні параметрів проблемної області, які окреслюють верхню і нижню межу всіх можливих поєднань корпусних характеристик.

Автентичність передбачає дотримання принципу оригінальності. У зв'язку з цим варто зазначити, що поняття автентичності часто досить широко інтерпретується. Так, зміст підручника може бути автентичним за своїм походженням, коли він написаний носіями мови, автентичним за своїми властивостями, коли написаний автором-носієм мови і не відрізняється від написаного в природно-мовному середовищі, і автентичним за своїми функціями, коли матеріал природно вписується в навчальний процес незалежно від авторства [17]. Тому головним критерієм автентичності тексту підручника слід вважати те, що його мовне наповнення не є штучно створеним з єдиною метою наповнення корпусу.

Відібраність ставить вимогу обмеження фактичного матеріалу шляхом відбору певних фрагментів мови з усього мовного континууму. Принцип відібраності заснований на розумінні того, що зміст корпусу не може бути випадковим і відбирається відповідно до визначених критеріїв. Це особливо важливо під час проведення вузькогалузевих досліджень. Наприклад, якщо метою є вивчення слів в академічній промові, то варто звертатися до корпусу, котрий складається тільки з фрагментів академічних текстів. Тому матеріалом нашого дослідження є виключно тексти найпопулярніших підручників китайської мови початкового рівня для іноземних студентів, а саме: «Новий практичний курс китайської мови» 《新实用汉语课本》, «Розвиваюча китайська мова. Комплексний курс. Початковий рівень»《发展汉语初级综合》, «Поглиблений курс китайської мови. Початковий рівень» 《博雅汉语初级》, «Курс китайської мови» (《汉语教程》).

Збалансованість полягає у введені до корпусу пропорційної кількості всієї лексики початкового рівня китайської мови, що дозволить отримати статистично достовірну інформацію про її використання. Тому корпус потребує необхідний і достатній обсяг текстів із кожного з відібраних підручників.

Машиночитаність передбачає спеціальну попередню підготовку текстів до їх подальшої комп'ютерної обробки, а саме оцифрування текстів з наступною їх автоматичною обробкою та анотуванням. Цей принцип також передбачає можливість автоматизованого аналізу корпусних даних програмними засобами. Для цього корпус повинен бути збережений в одному з електронних форматів (.txt, .rtf, .html та ін.) залежно від комп'ютерної програми, яка буде використовуватися для обробки текстових даних.

Принципи формування корпусу підручника китайської мови об'єднують низку вимог, виконання яких формує стратегію його побудови. Тому важливо знайти правильний баланс між повнотою явищ проблемної області, представленої в корпусі, та їх репрезентативністю, упорядкуванням текстових даних, форматами їх зберігання і наявністю програмних засобів їх обробки. Не менш важливо враховувати і принципи автентичності та відібраності.

Таким чином, згідно з принципами формування корпусу текстів, запропонованих В. Жуковською, ми трактуємо створений нами корпус текстів як «машиночитане, збалансоване, репрезентативне зібрання особливо розмічених (анотованих) текстів, відібраних згідно фіксованих параметрів для досягнення визначеної лінгвістичної мети» [9, с. 57].

Як уже зазначалося, процес створення корпусу пов'язаний також з проблемою виділення тексту, який слід включити до його складу. Ми дотримувались кількох змістовних принципів відбору тексту власне навчальних підручників, запропонованих А. Карамновим [10, с. 62]: принципу цілісності, принципу відповідності проблемній області, а також принципу структурної спрямованості.

Принцип **цілісності** полягає в орієнтації на певну категорію підручників, відповідно до вимог дослідження. Китайські підручники, наприклад, переважно складаються з певного набору тем, навколо яких будується вся навчальна діяльність – добираються завдання, формується зміст тощо. Мовне наповнення таких багаторівневих курсів (початковий, середній, високий) ґрунтується на прототиповому підході [16], коли з кожним рівнем розширюється лексичне поле тієї чи іншої теми. Таким чином, відповідно до принципу цілісності та залежно від поставлених завдань дослідження створюваний корпус підручника може включати як зміст книг всієї серії певного рівня (початкового, середнього та високого), так і матеріал окремих підручників. Крім того, може бути складений корпус окремих компонентів (робочого зошита, записів мультимедійних додатків тощо) або всього навчального комплекту.

Принцип **відповідності** проблемній області орієнтує укладача корпусу на визначення меж наявної мовної проблеми, яку повинен відобразити корпус. Відповідно до мети та завдань дослідження корпус підручника китайської мови може містити все його текстове наповнення, включаючи назву самого видання та уроків, посилання на словникові статті, граматичні пояснення тощо, або обмежуватися певною вибіркою даних. Така вибірка може проводитися, наприклад, за видами мовленнєвої діяльності, в результаті чого буде сформовано корпус завдань і текстів для читання, говоріння, аудіювання (включно з письмовим записом аудіо). Іншим критерієм, за яким відбирається матеріал, є його відповідність компонентам комунікативної компетенції (лінгвістичному, дискурсивному, прагматичному, стратегічному, культурному) [13, с. 67].

Принцип **структурної спрямованості** полягає в тому, що стратегія створення мовного корпусу підручника орієнтована на компонентний склад книги. Структура підручника не раз розглядалася в теорії і практиці навчання іноземних мов, зокрема в роботах Д. Зуєва [3], В. Бейлінсон [5], Е. Гельфмана, М. Холодної [6]. Згідно із зазначеним принципом може бути змодельований: а) корпус основного тексту підручника китайської мови, який буде містити матеріал, що є ключовим джерелом навчальної інформації; б) корпус додаткових текстів, який буде включати в себе матеріал, призначений для закріплення і поглиблення знань (наприклад, тексти для домашнього читання); в) корпус пояснювальних текстів, який об'єднує різні приклади, довідки, примітки, словникові статті та інший матеріал, покликаний забезпечити більш повне розуміння і засвоєння інформації. Крім цього, тут можна говорити і про побудову корпусу навчальних завдань (завдання і вправи), корпусу текстів для організації (заголовків, планів, параграфів, прикладів, нагадувань, підписів до зображень та схем тощо), корпусу текстів для орієнтування (передмови, вступи, зміст, рубрики, вказівки тощо).

Отже, змістові принципи створення корпусу підручника китайської мови, що підкреслюють ідею цілісності навчального матеріалу в багаторівневих і лінійних курсах, ідею відповідності корпусного змісту конкретній проблемі, а також ідею градуйованої значущості структурних компонентів видання, виконують роль орієнтирів у плануванні стратегії побудови мовного корпусу навчального підручника. Саме тому, згідно із зазначеними вище змістовими принципами створення корпусу, нами укладено корпус основних текстів підручників початкового рівня китайської мови для іноземців.

Технологічний процес створення власного корпус вимагав поступового виконання наступних кроків [9, с. 80]. Спершу були визначені джерела лінгвістичних даних. У нашому дослідженні – це публічно доступні тексти підручників з китайської мови початкового рівня. Наступний крок – введення даних. Існує три способи введення даних у корпус: адаптація даних в електронному форматі, сканування та ручне введення. У нашому дослідженні було використано метод ручного введення, оскільки нас цікавлять лише основні тексти кожного уроку підручників. У нашому дослідженні було враховано тексти вправ, пояснень граматичного матеріалу тощо. Ґрунтуючись на дослідженнях Ґо Шулуня (郭曙纶), ми вважаємо, що основні лексичні дані містяться в текстах, котрі «відкривають» урок та репрезентують ключову лексику, основні граматичні конструкції уроку і є базовим джерелом навчальної інформації [18, с. 116]. Після введення даних було здійснено анотування (розмітку) корпусів текстів, тобто процес введення формалізованої лінгвістичної інформації в електронний текст. Аплікативне призначення корпусних даних – морфологічні, синтаксичні, лексикологічні, лексикографічні дослідження – детермінує тип лінгвістичної анотації корпусу. Проф. Н. Дарчук надає необхідні для нашого дослідження пояснення типів лінгвістичної анотації корпусу. Зокрема, **морфологічна** анотація передбачає визначення морфологічних параметрів слова: частиномовну приналежність і категорійні ознаки кожної словоформи тексту. **Морфна** анотація полягає у сегментуванні кожної словоформи тексту за типами морфів і подальшому автоматичному укладанні серії морфемних та словотвірних словників із частотними характеристиками морфа/морфеми, за якими можна вивчати комбінаторно-дистрибутивну будову слова. **Синтаксична** анотація пов'язана з автоматичним опрацюванням кожного речення: виокремленням у ньому словосполучень та приписуванням кожному з них інформації про тип (дієслівний, іменниковий тощо), вид синтаксичного зв'язку та семантичного відношення. **Лексико-семантична** інформація, яка приписується кожному слову, відповідає таксономічній класифікації, розробленій відповідно для кожної частини мови [7, с. 67]. Слід зазначити, що у процесі обробки китайськомовних текстів за основу береться морфологічний аналіз. Важлива особливість китайської мови полягає в тому, що на письмі ієрогліфи не відокремлюються пробі-

Таблиця 1

Основний набір тегів частиномовної розмітки в корпусах китайської мови

№	Частина мови	Тег	Приклад
1.	Іменник	n	花/n ‘квіти’
2.	Іменник, що вказує на час	nt	星期一/nt ‘понеділок’
3.	Іменник, що вказує місцезнаходження	nd	门前/nl‘біля воріт’
7.	Географічна назва	ns	中国/ns‘Китай’
8.	Дієслово	v	跑/v‘бігати’
9.	Модальне дієслово	vu	能/vu‘могти’
10.	Прикметник	a	漂亮 /a‘красивий’
11.	Числівник	m	百/m‘сто’
12.	Рахівне слово	q	个/q‘універсальне рахівне слово’
13.	Прислівник	d	常常/d‘часто’
14.	Займенник	r	他/r‘він’
15.	Сполучник	c	但是/c‘проте’
17.	Вигук	e	啊/e‘о’
20.	Прийменник	p	由于/p‘через’
21.	Абревіатура	j	北大 （北京大学）/ j ‘Пекінський університет’
22.	Знак пунктуації	w	, /w

лом, що ускладнює поділ тексту на слова. Водночас через відсутність показників категорій числа, відмінка і роду, а також узгодження функція слова в китайській мові стає зрозумілою не на підставі форми слова, а завдяки його зв'язку з іншими словами [11, с. 35]. У процесі створення корпусу процедура поділу тексту на слова та словосполучення називається сегментацією, яка створює основу для розмітки.

У сучасній китайській мові слова розподіляються за двома групами: прості і складні [20, с. 86]. Прості слова складаються з однієї морфеми (один ієрогліф), наприклад: 钱 *гроші*, 人 *людина*. Складні слова утворюються двома і більше морфемами, між котрими існують різноманітні зв'язки, що ускладнюють сегментацію слів під час створення корпусу китайської мови. У 2006 році Інститут прикладної лінгвістики при Міністерстві освіти КНР запропонував «Принципи частиномовної розмітки під час обробки інформації китайською мовою», який встановлює конкретний стандарт морфологічної розмітки [19, с. 108]. Згідно з цим стандартом у китайській мові виділяються такі морфологічні категорії [12, с. 96]:

Окрім зазначених у таблиці 1, виділяють ще кілька важливих категорій.

У китайській мові існують атрибутивні слова, які відображають різницю між об'єктами одного типу або характеризують особливості чи якості предметів. Наприклад: 鸡 – загальна назва для курки і півня, відповідно 母鸡 *курка*, а 公鸡 *півень* – у цих словах жіночий рід виражається складоморфемою 母, а чоловічий рід – 公. Тому ієрогліфи 母 і 公 є атрибутивними складоморфемами і позначаються спеціальним тегом f: 公 / f鸡 /. Атрибутивні слова мають функцію, аналогічну функції прикметників. Однак, на відміну від прикметників, атрибутивні слова не вживаються самостійно у реченнях і перед ними не можна вживати прислівник [22, с. 75].

Перед деякими іменниками у китайських реченнях використовується додатковий член речення, який характеризує особливості або якості предметів. Наприклад:经理 – *директор*, 总经理 – *гендиректор*, 副经理 – *заступник директора*. У китайській мові ієрогліф 总 має значення *генеральний*, а 副 – *заступник*. Тим не менш, вони завжди стоять перед іменниками і окремо не вживаються. Наприклад: 书记 *секретар*, 总书记 *генеральний секретар*, 副书记 *заступник секретаря*. У китайській мові ці слова називаються препозитивними додатками, які позначаються тегом h – 副 / h 书记 / n.

Крім препозитивного додатку, в китайській мові існує постпозитивний додаток – ієрогліфи-суфікси, які утворюють нові слова, переважно – з іменниками. Наприклад, поєднання ієрогліфів 工程 *об'єкт* + 师 *майстер* = 工程师 *інженер*. Ці постпозитивні додатки позначаються тегом k – 药剂 / n 师 / k.

Згідно з «Принципами частиномовної розмітки при обробці інформації китайською мовою» [18, с. 117], під терміном «одиниця сегментації» розуміється основна одиниця, наділена певною семантичною або граматичною функцією при обробці китайськомовної інформації. Тобто китайські слова, внутрішня структура яких відповідає принципу композиційності, потребують поділу на одиниці сегментації. Якщо принцип композиційності не виконується, то в поділі немає необхідності. Принцип композиційності – значення складного висловлювання визначається значеннями його значущих компонентів плюс певний спосіб їх композиції [21, с. 21]. Суть композиційності полягає в тому, що значення складного висловлювання дорівнює поєднанню значень його складових частин. Так, значення слів 外语 *іноземна мова*, 洗手 *мити руки*, 头疼 *болить голова* формується за рахунок поєднання значень компонентів, а саме: 外语 *іноземна мова*, де 外 *іноземний*, а 语 *мова*; 洗手 *мити руки*, де 洗 *мити*, а 手 *рука*; 头疼 *болить голова*, де 头 *голова*, а 疼 *боліти*. Відповідно, внутрішня структура такого слова відповідає принципу композиційності і потребує поділу на одиниці сегментації, між якими ставиться скісна риска (слеш) з подальшим позначенням відповідними частиномовними тегами, як 新 / a 书 / n, 洗 / v 手 / n, 头 / n 疼 / v.

Як уже було сказано, слова в китайській мові представлені одним ієрогліфом, це прості слова, а поєднанням двох і більше ієрогліфів – складні слова. Очевидно, що прості слова не викликають труднощів під час сегментації: 钱 / n, 人 / n, 跑 / v. У складних словах між складоморфемами існують різноманітні зв'язки, які ускладнюють сегментацію під час створення корпусу текстів китайської мови.

Заключний етап – на основі вказаних вище принципів розмітки та сегментації тексту було здійснено конвертування розмічених текстів у структуру спеціалізованої лінгвістичної інформаційно-пошукової системи «Он-лайн корпус китайської мови – автоматична морфологічна розмітка та сегментація тексту» (分词和词性标注) [14] (рис. 1).

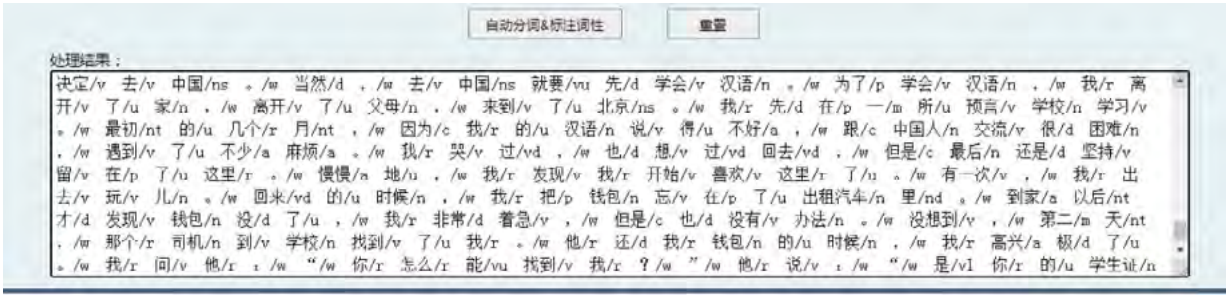

Рис 1. Результат автоматичної розмітки і сегментації корпусу (фрагмент)

Після проведення розмітки і сегментації текстів у корпусі за допомогою мережевої програми «Онлайн-корпус китайської мови – частотний аналіз тексту» (字词频率统计) [14] визначається частота слів. Слід зазначити, що інструментарій цієї програми дозволяє вводити і обробляти за один раз текст обсягом не більше 100 тисяч ієрогліфів. У результаті використання цієї програми отримано частотний список простих і складних слів, розташованих у порядку зменшення частоти. Наступними етапами дослідження є кількісний та якісний аналіз отриманих даних.

Висновки і пропозиції. На основі принципів формування корпусу текстів (репрезентативність, автентичність, відібраність, збалансованість, машиночитаність), змістових принципів відбору навчальних текстів (принцип цілісності, принцип відповідності проблемній області, принцип структурної спрямованості) та з урахуванням особливостей китайської мови вперше у вітчизняній науці здійснено спробу поетапного опису процесу створення корпусу текстів найпопулярніших підручників китайської мови початкового рівня. Перспективи подальших досліджень вбачаємо у кількісному дослідженні словникового складу вибраних комплектів підручників з метою виокремлення та системно-структурного аналізу базової лексики китайськомовного педагогічного дискурсу.

Список літератури:

1. Захаров В.П. Корпусная лингвистика: учеб. для студентов гуманитарных вузов. Иркутск : ИГЛУ, 2011. 161 с.

2. Зубов А.В. Корпусная лингвистика: возможности и проблемы. *Актуальные проблемы компьютерной лингвистики*. Минск. : МГЛУ, 2005. С. 71–72.

3. Зуев Д.Д. Школьный ученик. Москва : Педагогика, 1983. 240 с.

4. Баранов А.Н. Лингвистическая экспертиза текста: теория и практика: Учебное пособие. Москва : Флинта: Наука, 2007. 592 с.

5. Бейлинсон В.Г. Арсенал образования : характеристика, подготовка, конструирование учебных зданий. Москва : Книга, 1986. 218 с.

6. Гельфман Э., Холодная М. Психодидактика школьного учебника. Интеллектуальное воспитание учащихся. СПб. : Питер, 2007. 384 с.

7. Дарчук Н.П. Комп’ютерне анотування українського тексту: результати і перспективи. Київ : Освіта України. 2013. 339 с.

8. Драгалина-Черная Е.Г. Контекстуальность и композициональность. От принципа Фреге к когнитивным семантикам. Москва : МГЛУ, 2009. С. 66-84.

9. Жуковська В.В. Вступ до корпусної лінгвістики: навчальний посібник. Житомир : Вид-во ЖДУ ім. І. Франка, 2013. 140 с.

10. Карамнов А.С. Модель создания корпуса учебника английского язика. *Научный диалог*. Тамбов : Педагогика, 2013. С. 59–69.

11. Кочергин И.В. Очерки лингводидактики китайского язика. Москва : Восток – Запад, 2006. 190 с.

12. Лу Исинь. Гармонизация терминологии лингводидактики методами корпусной лингвистики: на материале русского и китайского языков: дис. канд. філол. наук. Санкт-Петербург. 2018. 242 с.

13. Мильруд Р.П. Введение в лингвистику. Учебное пособие для студентов педагогических вузов. Москва : Дрофа, 2005. 210 с.

14. Сайт Chinese Corpus online [Електронний ресурс] – Режим доступу до ресурсу: http://corpus.zhonghuayuwen.org/

15. Широков В.А. Корпусна лінгвістика. Київ : Довіра, 2005. 457 с.

16. Ghsoon R. English Coursebooks : Prototype Texts and Basic Vocabulary Norms. *ELT Journal*, 57, 2003. PP. 260-268.

17. Millrood R. Theory of Language Teaching : Linguistics, Didactics, Pedagogy. LAP LAMBERT Academic Publishing, 2010. 188 p.
18. 郭曙纶. 汉语语料库的建设及应用. 上海交通大学国际教育学院, 2016. 页 115–125.
19. 郭曙纶. 汉语语料库的建设及应用. 上海交通大学国际教育学院, 2018. 页 108–116.
20. 刘开瑛. 中文文本自动分词和标注. 商务印书馆, 2018. 页 85–96.
21. 张红武. 基于语料库的上海市初中语文教材语言统计与分析. 上海师范大学, 2007. 62页.
22. 周浪. 汉语术语词组组合模式. 南京理工大学, 2019. 页 74-92.

Sytnyk I. V. BUILDING TEXT CORPUS OF CHINESE ELEMENTARY TEXTBOOKS

This paper focuses specifically on the basic principles of building text corpus of Chinese elementary textbooks. The functions of text corpus in linguistics are presented. The specific features of Chinese language, which must be taken into account while building text corpus, are analyzed. The main stages of building text corpus of Chinese elementary textbooks are outlined. Five basic principles of building text corpus (representativeness, authenticity, selectivity, balance, machine readability) are characterized. It has been found that the process of building text corpus is also attached to the problem of selecting the text. The principles of selecting texts are analyzed: the principle of integrity, the principle of conformity to the problem area, the principle of structural orientation. Due to the fact that Chinese language is isolating language, a language in which each word form consists typically of a single morpheme, it has been defined that annotating of Chinese text corpus is made on the basis of morphological analysis. This paper also focuses on part-of-speech tagging and automatic text segmentation, the process of dividing written text into meaningful units. The part of speech tag-sets are presented and analyzed. Each step of building our own text corpus of Chinese elementary textbooks are outlined and described. Based on the analyzed principles of text corpus formation and text selection, as well as the main features of Chinese text segmentation and part-of-speech tagging, the selected texts of four the most popular Chinese elementary textbooks are converted into linguistic Internet program «Chinese Corpus online – Automatic POS-tagging and segmentation». The research results are presented. The first machine-readable, balanced, representative, marked (annotated) text corpus of Chinese elementary textbooks is built. These data thus need to be interpreted with caution and can be used in further quantitative and qualitative data analysis in order to provide a richer picture of Chinese elementary textbooks vocabulary phenomena under observation, and in particular, to be able to offer explanations.

Key words: *text, corpus, Chinese, textbooks, principles, elementary level.*