УДК 81.33

МЕТОДЫ СОЗДАНИЯ КИТАЙСКОГО КОРПУСА ТЕКСТОВ ЛИНГВОДИДАКТИКИ

Лу Исинь

Российский государственный педагогический университет им. А.И. Герцена, Санкт-Петербург

yixinhn@mail.ru

Поступила в редакцию 07.05.2017

Рассматривается проблема создания корпуса текстов китайского языка как специализированой поисковой системы для извлечения терминов из текстов в области обучения китайскому языку как иностранному и создания терминологических баз данных в данной области. Особое внимание уделяется основным этапам построения корпуса: отбор текстов на китайском языке в области обучения китайскому языку как иностранному, проведение разметки и сегментации текстов в корпусе, извлечение лексических единиц из корпуса и составление базового списка кандидатов в термины.

Ключевые слова: создание корпусов текстов, лингводидактика китайского языка, извлечение терминов.

Как и любая другая область науки, лингводидактика китайского языка нуждается в собственном терминологическом аппарате, необходимом для её эффективного развития и управления. Вместе с тем, в связи с бурным расцветом обучения китайскому языку как иностранному в России особенно актуальной становится проблема гармонизации терминологии лингводидактики в разных языках, предпосылкой для которой является создание терминологических баз данных в этой области.

Создание терминологических баз данных, как правило, опирается на анализ и оцифровывание уже опубликованных словарных источников и на результаты извлечения терминологии из корпусов текстов [1, с. 83]. Необходимо отметить, что на сегодняшний день терминосистема лингводидактики в китайском языке упорядочена крайне незначительно. Кроме того, следует учесть, что не было опубликовано ни одного словаря терминов данной области. Это является серьезным препятствием в развитии как самой лингводидактики, так и международного сотрудничества в этой предметной области. Таким образом, для извлечения терминов и создания терминологических баз данных в области лингводидактики китайского языка необходимо создание специализированного исследовательского корпуса текстов. Процесс его создания реализуется на 3 основных этапах:

1. Отбор текстов на китайском языке в области обучения китайскому языку как иностранному для создания корпуса;

2. Проведение разметки и сегментации текстов в корпусе и получение базовой лингвистической информации;

3. Извлечение лексических единиц (ЛЕ) из корпуса и составление базового списка кандидатов в термины.

Важной особенностью корпуса текстов является то, что он создается не просто как множество случайным образом объединенных текстов того или иного языка [2, с. 36]. При создании словаря на основе корпуса текстов необходимо определить принципы формирования выборочной совокупности для создания исследовательского корпуса текстов и ее необходимый и достаточный объем [3, с. 90]. Поскольку целью создания данного корпуса является извлечение терминов, отражающих терминополе лингводидактики китайского языка, и создание терминологической базы данных, то при создании корпуса текстов был установлен ряд специфических критериев отбора текстов:

1. Предметная ориентированность. Лингводидактический энциклопедический словарь определяет лингводидактику как «общую теорию обучения языку, включающую изложение теоретических основ такого обучения (представлений о содержании, целях и задачах, принципах, методах, процессе обучения) и его методических основ (обучение аспектам языка и видам речевой деятельности в конкретных условиях преподавания, организация учебного процесса, требования к профессии педагога)» [4, с. 140]. В нашем исследовании под китайской лингводидактикой понимается теория и методика обучения китайскому языку как иностранному, на этой основе в корпус включались только письменные научные тексты, посвященные теоретическим и методическим основам обучения китайскому языку как иностранному. При этом должны выбираться только тексты тех авторов, которые являются носителями китайского языка.

2. Соотнесенность по времени. В 1983 году в Китае создано научно-исследовательское общество по обучению китайскому языку как ино-

Таблица 1

Соотношение разных лексических единиц в китайском языке

Лексическая единица		Компонентный состав	Пример
слогоморфема		Один иероглиф	呢 (вопросительная частица)
Слова	простые слова	Одна слогоморфема	看(смотреть)
	сложные слова	Две слогоморфемы	电视(телевизор)
		Три слогоморфемы	老头儿(старик)
словосочетания		Два слова	中华/民族 (китайская нация)
		Три слова	视/听/教程 (аудиовизуальный курс)

странному, что официально ознаменовало рождение лингводидактики китайского языка как одной из отраслей науки. За последнее десятилетие благодаря повышению уровня экономического развития Китая и усилению его мощи, количество людей, изучающих китайский язык как иностранный, во всем мире значительно увеличилось. Вместе с тем теория и методика обучения в китайской лингводидактике беспрерывно совершенствуются. Поэтому в корпус отбираются только тексты, изданные после 2000 года.

3. Балансировка**.** Создаваемый корпус китайского языка должен обеспечить пропорциональное представление всех терминов лингводидактики, что позволит получать статистически достоверную информацию об их использовании. Поэтому корпус нуждается в необходимом и достаточном объеме текстов. Как показывают исследования [5, с. 6], у каждого слова китайского языка в среднем два значения, и каждое из них появляется в текстах минимум пять раз. Тогда корпус, создаваемый для составления словаря в 10 тысяч лексических единиц, должен включать 10 000×2×5=100 000 предложений. Если средняя длина предложения китайского языка – 25 иероглифов, то корпус должен быть объемом примерно в 4000 ЛЕ.

Русско-английский учебный словарь «Лингводидактика и тестирование» насчитывает примерно 1000 терминов, используемых при обучении русскому языку как иностранному [6]. Выдвинем гипотезу о том, что количество терминов лингводидактики в китайском языке приблизительно равно количеству терминов русского языка. Согласно приведенным ранее рассуждениям корпус, предназначенный для извлечения подобного объема терминов, должен включать примерно 400 тысяч ЛЕ – иероглифов.

Согласно всем выше перечисленным критериям для создания корпуса были выбраны следующие издания: «对外汉语教学概论 (Введение в обучение китайскому языку как иностранному)», издано в 2004 году в Пекине под редакцией Чен Чжантань и Ю Геньюань объемом в 304 756 иероглифов [7], и «汉语可以这样教—语言技能篇 (Преподавание китайского языка: языковой навык)», издано в 2016 году в Пекине под редакцией Чжао Циминь объемом в 95 479 иероглифов [8]. Таким образом, данный корпус содержит 400 235 иероглифа. Эти две книги очень популярны в области лингводидактики китайского языка в Китае и используются в качестве учебных пособий для подготовки преподавателей китайского языка как иностранного.

Специальная предварительная подготовка текстов к их последующей компьютерной обработке представляет собой оцифровывание текстов с последующей их вычиткой и расположением на магнитном носителе. Следует отметить, что при введении текстов в компьютер для статистически достоверных результатов необходимо придерживаться принципа оригинальности, то есть с уважением относиться к первоисточнику, не изменяя его.

Как звуковое, смысловое, интонационное и графическое единство в китайском языке иероглиф представляет собой слогоморфему, большинство слогоморфем имеет собственные значения и может использоваться самостоятельно, в то же время некоторые слогоморфемы обладают только грамматическими функциями и отдельно не употребляются. В современном китайском языке слова делятся на две группы: простые и сложные. Простые слова состоят из одной слогоморфемы (один иероглиф). Сложные слова образуются двумя или тремя слогоморфемами. Как показывают исследования [5, с. 24], отношения между слогоморфемами, словами и словосочетаниями можно представить следующим образом (см. табл. 1).

Сложной задачей, решаемой при создании корпусов текстов для языков, графика которых отлична от латиницы и кириллицы, является не только их оцифровывание, но и выделение границ слов. В письменной форме китайского языка между иероглифами (слогоморфемами) от-

Таблица 2

Основной набор тэгов частеречной разметки в корпусах китайского языка

Часть речи	Тэг	Часть речи	Тэг
Имя существительное	n	Прилагательное	a
Существительное времени	nt	Атрибутивные слогоморфемы	f
Существительное, означающее азимутальное направление	nd	Числительное	m
Существительное, указывающее местонахождение	nl	Счётное слово	q
Имя собственное	nh	Наречие	d
Фамилия	nhf	Местоимение	r
Имя	nhg	Союз	c
Географическое название	ns	Частица	u
Названия учреждений, организаций и компаний	ni	Междометие	e
Глагол	v	Звукоподражание	o
Глагол направленного	vd	Идиома	i
Глагол, выполняющий функцию связки	vl	Предлог	p
Модальный глагол	vu	Аббревиатура	j
Суффикс	k	Префикс	h
Пунктуационный знак	w		

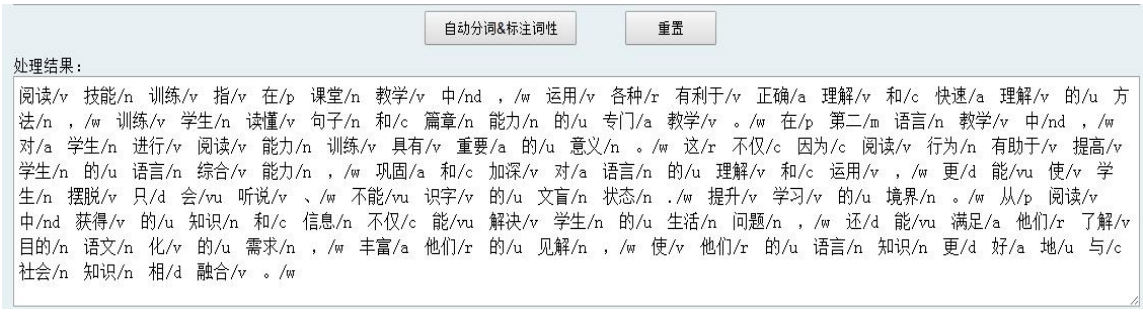

Рис. 1. Автоматическая разметка и сегментация корпуса с помощью программы *Automatic POS-tagging and segmentation*

сутствуют пробелы, т. е. потенциальные границы между единицами китайского языка существуют, но не указываются на письме. В связи с этим возникает необходимость разбить поток иероглифов в тексте на естественном языке на отдельные значимые единицы — слова, т. е. провести сегментацию. Частеречная разметка (тэгирование) уже сегментированного потока слогоморфем и сама сегментация представляют основу для дальнейшего исследования (см. табл. 2) [9, с. 25].

На основе указанных выше принципов разметка и сегментация текстов в корпусе проводилась при помощи сетевой программы *Chinese Corpus online — Automatic POS-tagging and segmentation* [10], предназначенной для лингвостатистического анализа текста (рис. 1).

После проведения разметки и сегментации текстов в корпусе при помощи сетевой программы *Chinese Corpus online — Frequency statistics of words* [10] определяется частота слов. Следует отметить, что инструментарий данной программы позволяет вводить и обрабатывать за один раз текст объемом не больше 100 тысяч иероглифов. Поэтому созданный корпус объемом в 400 тысяч нуждается в разделении на 4 части. В результате использования программы получены 4 словаря простых и сложных слов, который включает не только кандидаты в термины, но и общеупотребительные ЛЕ. Эти ЛЕ (стоп-слова) необходимо удалить из полученных частотных словарей.

Для формирования списка стоп-слов необходимо: а) извлечь из Интернета [11] список китайских стоп-слов, общих для разных предметных областей; б) на основе изучения текстов опубликованных научных статей, которые были отобраны из текстов официального веб-сайта CNKI (China National Knowledge Infrastructure) [12] в области китайской лингводидактики, вручную определить специальные стоп-слова, характерные только для данной области. На основе этих двух групп ЛЕ был составлен список стоп-слов, которые были удалены из полученных четырех частотных словарей. Для дальнейшего исследования после фильтрации все полученные словари были объединены в таблицу, фрагмент которой представлен ниже (табл. 3).

Китайский ученый Фэнь Жиуэн исследовал способ словообразования терминов, различаю-

Таблица 3

Частотный список слов (фрагмент)

Слово	Частота в первой части	Частота во второй части	Частота в третьей части	Частота в четвертой части	Суммарная частота
阅读 (чтение)	19	8	5	146	178
交流 (коммуникация)	49	72	45	8	174
重点 (ключ)	51	8	26	80	165
文章 (статья)	52	85	25	0	162
声调 (тон)	0	71	87	0	158

Таблица 4

Способы словообразования китайских многокомпонентных терминов

Длина термина	Способ словообразования	Пример
2 слова	n/v + n/v	表达/v能力/n (способность выражения)
	a + nv	第二/ a语言/n (второй язык)
	f + n	高级/ f教程/ n (продвинутый курс)
	m + n	四/ m声/ n (четыре тона)
3 слова	n/v + n/v + n/v	汉语/n水平/n考试/n (стандартизированный квалификационный экзамен по китайскому языку)
	a + n/v + n	第二/ a语言/n教学/ n (обучение второму языку)
	d + v + n	常/d用/v词/n (обиходное слово)
	f + v + n	初级/f口译/v教程/n (первоначальный курс устного перевода)

Обозначения: a – прилагательное, f – атрибутивные слогоморфемы, c – союз, d – наречие, m – числительное, n – существительное, v – глагол, u – вспомогательные слова, где n/v – омоним существительное/глагол.

щихся длиной [13], и пришел к выводу, что существительные и глаголы чаще могут быть либо однословными терминами, составленными из одиночных простых или сложных слов, либо основами (ядрами) многокомпонентных терминов, составленными из нескольких слов. Поэтому для дальнейшего исследования из табл. 3 были отобраны все существительные и глаголы, которые и составили список кандидатов в однословные термины.

Дальнейшая работа заключается в определении степени терминологичности кандидатов в термины и установлении списка реальных терминов. Табл. 3 показывает, что высокочастотные слова появляются во всех частях корпуса, а некоторые слова появляются только в одной или двух частях корпуса. Слова, встретившиеся во всех частях корпуса, рассматриваются как реальные однословные термины в области лингводидактики, и из них создается первый список однословных терминов. Остальные слова нуждаются в определении степени терминологичности (*termhood*). Одним из методов оценки степени терминологичности является не зависящий от предметной области метод автоматического выявления терминов в тексте, позволяющий упорядочивать их по степени терминологичности, которую принято называть C-Value. В работе Баррона-Кедено [14] C-Value обобщается на случай однословных терминов путем добавления константы к логарифму:

$$C\text{-}Value(t) = \begin{cases} c(t)TF(t), \text{если } \{s : t \subset s\} \\ c(t)(TF(t) - \dfrac{\sum_s TF(s)}{|\{s : t \subset s\}|}), \text{иначе} \end{cases}$$

где TF — частота вхождений кандидата в термины, $c(t)= i +\log_2|t|$. Автор отмечает, что изначально пробовал значение $i = 0.1$ для того, чтобы вносить меньше искажений в исходную формулу, однако в ходе экспериментов обнаружил, что наибольшую эффективность показывает значение $i = 1$.

На основе указанной выше формулы кандидаты в однословные термины упорядочиваются по степени их терминологичности. В то же время можно вычислять среднее значение в списке кандидатов в термины как пороговое значение и извлекать для окончательной проверки только те кандидаты в термины, у которых мера C-Value выше определенного порога. Данный список терминов объединяется с первым списком однословных терминов. Так получается окончательный список однословных терминов.

Как отмечает Фэнь Жиуэн [12], китайские многокомпонентные термины обычно состоят из цепочек длиной от 2 до 3 слов, и существуют следующие способы их образования (см. табл. 4).

Таблица 5

Многокомпонентные термины с ядерным словом 语音

Двусловные термины с ядерным словом *语音*	Трехсловные термины с ядерным словом *语音*
语音/n教学/v (обучение фонетике)	汉语/n语音/n辨析/v (распознавание китайской фонетики)
语音/n能力/n (фонетический навык)	汉语/n语音/n习得/v (овладение китайской фонетикой)
语音/n训练/v (тренировка по фонетике)	汉语/n语音/n理论/ n (теория китайской фонетики)

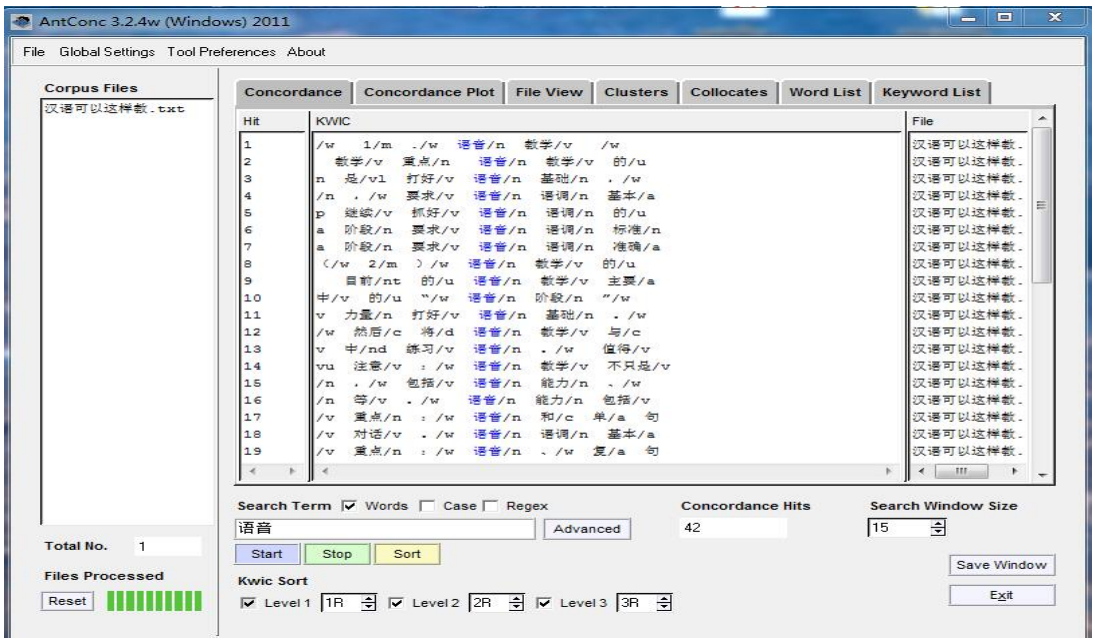

Рис. 2. Пример результатов работы программы AntConc-Concordance

Извлечение многокомпонетных терминов может производиться при помощи сетевой программы AntConc на основе поиска устойчивых словосочетаний с установленными ранее однословными терминами в качестве ядер. Инструмент Concordance программы AntConc позволяет рассматривать окружение выявленных терминов. Табл. 4 показывает, что китайские многокомпонентные термины состоят из двух или трех слов. Программа AntConc-конкордансер может обработать только размеченный корпус текстов на китайском языке, и в процессе работы тэгированная последовательность иероглифов-слогоморфем рассматривается как одно слово. Поэтому для того чтобы не пропустить цепочки слогоморфем, которые потенциально могут быть многокомпонентными терминами, в программе Context Horizon с левой и правой сторон указывается параметр 6, то есть отбираются по 6 иероглифов с обеих сторон ядра. Словосочетания, компоненты которых связаны между собой и отвечают вышеуказанным условиям (см. табл. 4), извлекаются как многословные термины.

Рассмотрим в качестве примера термин **语音** (фонетика), программой Context Horizon выделены 42 цепочки с данным ядерным словом (рис. 2).

Далее извлекаются те многокомпонентные термины с ядерным словом, чья внутренняя структура соответствует условиям, показанным в табл. 4 (см. табл. 5).

В результате подобного анализа все извлеченные однословные и многокомпонентные термины объединяются для ручной проверки и составления окончательного списка терминов. Таким образом, можно утверждать, что последовательное применение мер лингвистического и количественного анализа к специализированному корпусу текстов позволяет создать список кандидатов в термины, резко сокращающий работу терминолога и позволяющий создавать реальные глоссарии предметной области.

Список литературы

1. Беляева Л.Н. [и др.] Лексикографический потенциал современных лингвистических технологий. СПб.: Книжный дом, 2014. 168 с.

2. Захаров В.П., Богданова С.Ю. Корпусная лингвистика. Иркутск: ИГЛУ, 2011. 161 с.

3. Беляева Л.Н. Корпусная лингвистика и перевод: потенциал и ограничения // Труды Международной конференции «Корпусная лингвистика – 2011» / Филол. фак. СпбГУ. СПб., С. 87–91.

4. Щукин А.Н. Лингводидактический энциклопедический словарь: более 2000 единиц. М.: Астрель, 2007. 746 с.

5. Guo Shulun. The Construction and Application of Chinese Corpus [M]. Shanghai: Shanghai Foreign Language Education Press, 2012.

6. Беляева Л.Н. [и др.] Лингводидактика и тестирование: англо-русский и русско-английский учебный словарь. СПб.: Книжный дом, 2014. 110 с.

7. Chen Zhangtai, Yu Genyuan. An Introduction to Teaching Chinese as a Foreign Language [M]. Beijing: Commercial Press, 2004.

8. Zhao Jinming. Teaching Chinese as Foreign Language – Language Skill [M]. Beijing: Commercial Press, 2016.

9. Лу Исинь. Принципы создания корпусов китайского языка // Известия РГПУ им. А.И. Герцена. 2016. № 181. С. 22–29.

10. Chinese Corpus online [Электронный ресурс]. Режим доступа: http://cncor pus.org

11. Chinese Stoplist [EB/OL] [2012.11.20] [Электронный ресурс]. Режим доступа: http: //www.smart peer.net/myfiles/stopwords.utf8.txt).

12. China National Knowledge Infrastructure [Электронный ресурс]. Режим доступа: www.cnki.net.

13. Feng Zhiwei. An Introduction to Modern Terminology [M]. Beijing: Language & Culture Press, 1999.

14. Barron-Cedeno A., Sierra G., Drouin P. et al. An Improved Automatic Term Recognition Method for Spanish // Computational Linguistics and Intelligent Text Processing. Springer, 2009. P. 125–136.

THE METHODS OF BUILDING CORPORA FOR CHINESE EDUCATIONAL LINGUISTICS

Yixin Lu

The paper deals with the problem of building a Chinese corpus as a specialized search system for extracting terms from texts in the field of teaching Chinese as a foreign language. The corpus also serves for building a terminology database. This process of building the corpus is implemented in 3 main steps: selection of Chinese texts in the field of teaching Chinese as a foreign language for building a corpus; segmentation and POS (Part-of-speech) tagging of text words and providing basic linguistic information; extracting terms from the corpus and compiling a list of terms.

Keywords: building text corpora, Chinese educational linguistics, term extraction.

References

1. Belyaeva L.N. [i dr.] Leksikograficheskij potencial sovremennyh lingvisticheskih tekhnologij. SPb.: Knizhnyj dom, 2014. 168 s.

2. Zaharov V.P., Bogdanova S.Yu. Korpusnaya lingvistika. Irkutsk: IGLU, 2011. 161 s.

3. Belyaeva L.N. Korpusnaya lingvistika i perevod: potencial i ogranicheniya // Trudy Mezhdunarodnoj konferencii «Korpusnaya lingvistika – 2011» / Filol. fak. SPbGU. SPb., S. 87–91.

4. Shchukin A.N. Lingvodidakticheskij ehnciklopedicheskij slovar': bolee 2000 edinic. M.: Astrel', 2007. 746 s.

5. Guo Shulun. The Sonstruction and Application of Chinese Corpus [M]. Shanghai: Shanghai Foreign Language Education Press, 2012.

6. Belyaeva L.N. [i dr.] Lingvodidaktika i testirovanie: anglo-russkij i russko-anglijskij uchebnyj slovar'. SPb.: Knizhnyj dom, 2014. 110 s.

7. Chen Zhangtai, Yu Genyuan. An Introduction to Teaching Chinese as a Foreign Language [M]. Beijing: Commercial Press, 2004.

8. Zhao Jinming. Teaching Chinese as Foreign Language – Language Skill [M]. Beijing: Commercial Press, 2016.

9. Lu Isin'. Principy sozdaniya korpusov kitajskogo yazyka // Izvestiya RGPU im. A.I. Gercena. 2016. № 181. S. 22–29.

10. Chinese Corpus online [Ehlektronnyj resurs]. Rezhim dostupa: http://cncor pus.org

11. Chinese Stoplist [EB/OL] [2012.11.20] [Ehlektronnyj resurs]. Rezhim dostupa: http://www.smartpeer.net/myfiles/stopwords.utf8.txt).

12. China National Knowledge Infrastructure [Ehlektronnyj resurs]. Rezhim dostupa: www.cnki.net.

13. Feng Zhiwei. An Introduction to Modern Terminology [M]. Beijing: Language & Culture Press, 1999.

14. Barron-Cedeno A., Sierra G., Drouin P. et al. An Improved Automatic Term Recognition Method for Spanish // Computational Linguistics and Intelligent Text Processing. Springer, 2009. P. 125–136.