УДК 811.581.81’332

ПРИНЦИПЫ ИЗВЛЕЧЕНИЯ ЛИНГВОДИДАКТИЧЕСКИХ ТЕРМИНОВ ИЗ КОРПУСОВ КИТАЙСКИХ ТЕКСТОВ

Лу Исинь

Российский государственный педагогический университет имени А. И. Герцена

Поступила в редакцию 9 декабря 2016 г.

Аннотация: *статья посвящена проблеме создания терминологических баз данных в области китайской лингводидактики. Основное внимание уделено выяснению и описанию принципов и методов извлечения терминов из корпусов текстов в области обучения китайскому языку как иностранному.*
Ключевые слова: *извлечение терминов, китайская лингводидактика, корпус.*

Abstract: *the paper deals with building a terminology database for Chinese educational linguistics. The emphasis is on the elucidation and description of principles and methods for corpus terms extraction, the corpus includes texts from the field of teaching Chinese as a foreign language.*
Key words: *term extraction, Chinese educational linguistics, corpus.*

Создание терминологических баз данных, как правило, основано на анализе и оцифровке уже опубликованных словарных источников, а также на результатах извлечения терминологии из корпусов текстов [1, с. 83].

Предмет исследования данной работы – принципы и методы извлечения терминов из размеченного корпуса текстов в области китайской лингводидактики (обучения китайскому языку как иностранному).

Педагогический 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 о методических основах обучения языку (обучение аспектам языка и видам речевой деятельности в конкретных условиях преподавания, организация учебного процесса, требования к профессии педагога)». В данной статье под китайской лингводидактикой понимается теория и методика обучения китайскому языку как иностранному [2, с. 140].

Лингводидактический энциклопедический словарь определяет лингводидактические термины как «одну или несколько лексических единиц в комбинации, предельно полно и точно выражающих конкретное лингводидактическое понятие, подчиняющихся общелингвистическим законам словообразования и взаимодействиям лексических единиц в синтаксических единствах и уточняющих свое значение в контексте данного специального текста по методике преподавания иностранных языков» [3, с. 245].

В данной статье извлечение терминов (как универбов, так и многокомпонентных лексических единиц – коллокаций) основано на объединении статистических оценок и семантического анализа. Этот процесс реализуется на трех основных этапах:

1. Создание корпуса. Этот этап состоит из отбора текстов в области лингводидактики, создания корпуса на китайском языке, проведения частеречной разметки и сегментации текстов в корпусе и получения частотных словарей [4].

2. Выбор кандидатов в термины. Данный этап состоит в фильтрации слов и словосочетаний, извлеченных из корпуса текстов, по статистическим и лингвистическим критериям для выделения цепочек слогоморфем, которые потенциально могут быть терминологическими единицами.

3. Анализ выбранных кандидатов в термины. Этот этап состоит в применении ряда процедур выбора терминов, которые должны оценить меру терминологичности кандидатов в термины [1, с. 94]. Используя меру C-Value [5; 6], определяется степень терминологичности кандидатов в термины в области китайской лингводидактики.

Рассмотрим особенности организации работы на втором и третьем этапе.

Второй этап, в свою очередь, можно разделить на следующие две фазы.

I. Как отмечает Ли Юн [7, p. 35] (ср. [8]), китайские термины обычно состоят из цепочек длиной от 2 до 6 иероглифов, которые составляют 76,9 % от общего числа терминов. В то же время термины, состоящие из 4 иероглифов, составляют 26,1813 % от общего числа терминов; термины, образованные 5 иероглифами, составляют 18,9775 % от общего числа; термины из 6 иероглифов составляют 17,8932 %.

Слова в китайском языке представлены отдельными иероглифами (однослоги) и сочетаниями двух

и более иероглифов (двуслоги и многослоги). Китайский ученый Фэнь Жиуэн также исследовал способ словообразования терминов, различающихся длинами [9], и пришел к выводу (табл. 1), что существительные и глаголы чаще могут быть однословными терминами или основами многословных терминов.

Таблица 1

Способ словообразования китайских терминов

Длина терминов	Способ словообразования	Пример
1 слово	n,v	阅读/ n (чтение), 说/ v (говорить)
2 слова	nv+nv, a+nv, b+n, m+n	表达/v能力/n (способность выражения)
3 слова	nv+nv+nv, a+nv+n, d+v+n, b+v+n	第二/ a语言/n教学/ n (обучение второму языку)

Примечание. В таблице приняты обозначения: a – прилагательное, b – атрибутивные слогоморфемы, c – союз, d – наречие, m – числительное, n – существительное, v – глагол, u – вспомогательные слова, где nv – существительное или глагол.

Таким образом, для дальнейшего исследования корпуса из частотного словаря были отобраны все существительные и глаголы, которые и составили первый вариант списка кандидатов в термины.

В целях снижения шума была проведена фильтрация, заключающаяся в удалении из массива кандидатов в термины стоп-слов из заранее составленного списка. Для формирования такого списка необходимо: а) извлечь из системы Интернет [10] список китайских стоп-слов, общих для разных предметных областей; б) на основе изучения текстов авторефератов в области китайской лингводидактики вручную определить специальные стоп-слова, характерные только для данной области. На основе этих двух списков производится удаление общих и специальных стоп-слов, которые встречаются в списке выявленных существительных и глаголов. После этого получаем список кандидатов в однословные термины.

II. Извлечение кандидатов в многокомпонентные термины происходит при помощи сетевой программы AntConc-конкордансера, предназначенного для лингвостатистического анализа текста [11]. При помощи сетевой программы AntConc можно получить данные об окружении лексических единиц, используя такой сервисный инструмент, как конкорданс.

Инструмент Concordance программы AntConc позволяет рассматривать окружение выявленных существительных или глаголов. Как было указано выше, максимальное количество иероглифов в китайских терминах равно 6, в программе Context Horizon с левой и правой сторон указывается 6, т. е. отбирается по 6 иероглифов с обеих сторон ядра. Частота появления цепочки лексических единиц с ядрами регистрируется. Словосочетания, компоненты которых связаны между собой и отвечают вышеуказанным условиям (см. табл. 1), извлекаются как кандидаты в многословные термины. В результате получаем два списка кандидатов в термины: список двусловных кандидатов и список кандидатов, состоящих из трех слов.

Все полученные списки объединяются в одну таблицу, которая представлена ниже (табл. 2).

Таблица 2

Списки кандидатов в термины (фрагмент)

Кандидаты в однословные термины	Кандидаты в двусловные термины	Кандидаты в трехсловные термины
语言/n (язык)	第二/a语言/n (второй язык)	第二/a语言/n教学/n (обучение по второму языку)
汉语/n (китайский язык)	对外/b汉语/n (китайский язык как иностранный)	对外/b汉语/n教学/n (обучение китайскому языку как иностранному)
表达/v (выразить)	口头/a表达/v (вербальная фомулировка)	口头/a表达/v能力/n (способность вербальной фомулировки)
交际/n (коммуникация)	跨文化/a交际/n (международная коммуникация)	语言/n交际/n活动/n (языковые коммуникативные деятельности)

Третий этап заключается в определении степени терминологичности кандидатов в термины и установлении списка реальных терминов. Дело в том, что выделенный на предыдущих этапах анализа список кандидатов в термины обычно включает в себя не только термины, общеупотребительные коллокации, но и бессмысленные цепочки лексических единиц.

Большинство автоматизированных систем извлечения терминов используют либо статистический, либо лингвистический подход. В последнее время появились гибридные подходы, использование которых представляет собой попытку преодоления ограничений односторонних подходов к решению задачи извлечения терминов на основе как лингвистических, так и статистических элементов [1, с. 90]. 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.

Мера C-Value вычисляется непосредственно на основе характеристик цепочек лексических единиц, являющихся кандидатами в термины. К ним относятся:

1. Суммарная частота цепочки выбранных лексические единицы в корпусе текстов.

2. Частота цепочки выбранных лексических единиц, встретившейся как части других более длинных выбранных цепочек.

3. Количество таких более длинных кандидатов в термины.

4. Длина цепочки кандидатов в термины [5, с. 218].

В исследовании [12] предложена следующая формула вычисления C-Value:

$$C\text{-}Value(t) = \begin{cases} \log_2|t| \cdot f(t), & \text{если } \{s : t \subset s\} = \varnothing; \\ \log_2|t| \cdot \left(f(t) - \dfrac{\sum_s f(s)}{|\{s : t \subset s\}|} \right), & \text{иначе}, \{s : t \subset s\} = \varnothing \end{cases}$$

где t – кандидат в термины; $|t|$ – длина кандидата t (в словах); $f(t)$ – частота вхождений t в коллекции текстов; s – множество кандидатов, окружающих кандидата t, т. е. таких кандидатов, что t является их подстрокой.

Важно отметить, что мера C-Value предназначена для извлечения только многословных терминов: иначе выражение под логарифмом обнуляет значение признака.

В работе Баррона-Кедено [13] C-Value обобщается на случай однословных терминов путем добавления константы к логарифму:

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

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 $i = 1$.

При помощи указанных выше двух формул, кандидаты в однословные и многословные термины упорядочиваются по степени их терминологичности, при этом можно просматривать списки сверху вниз. В то же время можно вычислять среднее значение в каждом 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.

В результате такого анализа все извлеченные кандидаты в термины объединяются для ручной проверки и получения 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.

ЛИТЕРАТУРА

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

2. Педагогический энциклопедический словарь. – М. : Большая российская энциклопедия, 2003. – 528 с.

3. *Щукин А. Н.* Лингводидактический энциклопедический словарь : более 2000 единиц / А. Н. Щукин. – М. : Астрель, 2007. – 746 с.

4. *Лу Исинь*. Принципы создания корпусов китайского языка / Исинь Лу // Известия РГПУ им. А. И. Герцена. – 2016. – № 181. – С. 22–19.

5. *Беляева Л. Н.* Системы и процедуры выделения терминов из текстов / Л. Н. Беляева // Терминология и знание : материалы IV Междунар. симпозиума. – М. : Вест-Консалтинг, 2014. – С. 212–223.

6. *Li Chao.* Exploiting Domain Interdependence for Multi – Word Terms Extraction [J] / Chao Li // Journal of Chinese Information Processing. – 2010. – № 24 (1). P. 94–98.

7. *Li Yun.* Automatic Term Extraction in the Field of Information Technology [J] / Yun Li // Terminology Standardization and Information Technology. – 2003. – № 1. P. 32–37.

8. *Zhang Pu.* Structural Featurs and Distributions of Chinese Terms in the Corpus from Information Field [J] / Zhang Pu // New Technology of Library and Information Service. – 2007. – № 12. – P. 54–58.

9. *Feng Zhiwei.* An Introduction to Modern Terminology [M] / Zhiwei Feng. – Beijing, Language & Culture Press, 1999.

10. Chinese Stoplist [EB/OL]. [2012–11–20]. – Mode of access: http: // www. smartpeer. net/myfiles/stopwords – utf8.txt

11. *Беляева Л. Н.* Сетевой инструментарий филолога / Л. Н. Беляева, К. Р. Пиотровская. – СПб. : Книжный дом, 2014. – 48 с.

12. *Frantzi K.* Automatic recognition of multi-word terms. The c-value/nc-value method / K. Frantzi, S. Ananiadu, H. Mima // Internationl Journal on Digital Libraries. – 2000. – Vol. 3, № 2. – P. 115–130.

13. *Barron-Cedeno A.* An improved automatic term recognition method for Spanish / A. Barron-Cedeno [et al.] // Computational Linguistics and Intelligent Text Processing. – Springer, 2009. – P. 125–136.

Российский государственный педагогический университет имени А. И. Герцена

Лу Исинь, аспирант кафедры немецкой филологии

E-mail: yixinhn@mail.ru

Тел.: 8-911-285-23-72

Russian State Pedagogical University named after A. I. Herzen

Lu Yixin, Post-graduate Student of the German Philology Department

E-mail: yixinhn@mail.ru

Tel.: 8-911-285-23-72