Е.Н.Колпачкова

E.N.Kolpachkova

КОРПУСЫ КИТАЙСКОГО ЯЗЫКА: СОВРЕМЕННОЕ СОСТОЯНИЕ И ОСНОВНЫЕ ПРОБЛЕМЫ
CHINESE LANGUAGE CORPORA: AN OVERVIEW AND MAJOR PROBLEMS

Аннотация. В статье дается обзор корпусов китайского языка с описанием их основных особенностей, а также обсуждаются проблемы аннотирования корпусов на языках с бедной морфологией, каким является китайский язык. Объективные сложности связаны с неоднозначностью грамматической классификации словарного состава языка и затрудненным сегментированием текста на отдельные словоформы.

Ключевые слова: китайский язык, корпус, корпусная лингвистика, разметка, части речи.

Abstract. The article provides an up-to-date survey of the major Chinese language corpora available so far. It begins with a brief overview of the existing corpora in mainland China and abroad, then we discuss some difficulties in the Chinese corpus-building process. Since two main problems of the Chinese grammar - word segmentation and part-of-speech tagging- haven't been solved yet, corpus annotation and parsing tools are still under development.

Keywords: chinese language, corpus linguistics, annotation, POS-tagging.

1. Введение

Исследования по разработке машинного фонда текстов на китайском языке начались в 1982 году, когда появился первый в

Китае корпус английского языка JDEST, и с тех пор корпусно-ориентированные изыскания активно ведутся научными группами в почти двух десятках университетов и институтов в КНР, а также в Гонконге, на Тайване и за рубежом. В результате этой деятельности за три десятилетия, что насчитывает история современной корпусной лингвистики в Китае, было создано свыше двух десятков корпусов китайского языка, а также целый ряд двуязычных корпусов.

2. Современные корпусы китайского языка

В настоящее время продолжает реализовываться целый ряд проектов по созданию общих и специализированных корпусов китайского языка. Самым ранним из общедоступных явился Корпус современного китайского языка (The Modern Chinese Language Corpus) Центра китайской лингвистики при Пекинском университете[1]. Метод поиска в данном корпусе базируется на фактическом расстоянии между иероглифами/слогами и позволяет построить конкорданс – список всех употреблений данной словоформы в контексте. В Корпусе современного китайского языка имеется только метаразметка, при этом отсутствуют такие обязательные признаки корпуса, как морфологическая и синтаксическая разметки, поэтому в строгом смысле эту базу нужно рассматривать не как корпус языка, а скорее как текстовый архив. Неаннотированность делает использование данного корпуса в области грамматических исследований довольно затруднительным.

Значительным шагом вперед в корпусной лингвистике КНР стало создание Сбалансированного корпуса китайского языка[2] ,

[1] http://ccl.pku.edu.cn:8080/ccl_corpus/index.jsp

[2] www.cncorpus.org

который представлен не простым текстовым форматом, а снабжен лингвистической разметкой, позволяющей осуществлять поиск по морфологическим признакам словоформ. Этот сбалансированный представительный корпус, созданный по заказу Государственного Комитета по работе в области языка и письменности КНР, имеет два подкорпуса - почти 13 млн. словоупотреблений современного языка и около 100 млн. иероглифов для корпуса древних текстов.

Частеречная разметка этого корпуса включает не только указание на часть речи, но и признаки грамматических категорий, свойственных в китайском языке данному лексико-грамматическому классу. По замыслу создателей, схема морфологической разметки предполагает набор тэгов для 13 классов слов первого уровня (一级) и 16 подклассов （二级）, отдельная разметка предусмотрена для идиом, аббревиатур и отдельных морфем. Предусмотрена опция поиска по «сырому» тексту, также доступна выдача данных в формате KWIC (key word in context) с выравниванием ключевых слов по центру. Древнекитайский корпус не размечен, поиск возможен по словам, словосочетаниям, с возможной выдачей результатов в формате KWIC. На нынешнем этапе ведется постоянное пополнение и усовершенствование функциональности данного корпуса, являющегося сегодня, пожалуй, самым современным и функциональным из всех корпусов китайского языка.

В течение последних десятилетий во многих университетах и научных центрах по всему миру ведется работа над созданием корпусов текстов на китайском языке, пионерами которых были создатели китайского варианта Корпуса Лидсского университета[3].

[3] http://corpus.leeds.ac.uk/internet.html

Для нужд как грамматических, так и лексикографических исследований определенный интерес может представлять аннотировнный корпус текстов Academia Sinica Balanced Corpus of Modern Chinese (или Sinica Corpus)[4], имеющий морфологическую и синтаксическую разметку в соответствии с принятыми на Тайване стандартами сегментации слов и выделения частей речи.

Наиболее интенсивно идет строительство корпусов в самом Китае, здесь активно создаются и специализированные текстовые базы, например, Корпус прессы LIVAC[5] Гонконгского педагогического института, аннотированный Корпус газеты Жэньминь жибао (北京大学《人民日报》标注语料库)[6], Речевой корпус современного пекинского диалекта (北京地区现场即席话语语料库) [7] и многие другие корпусы, создаваемые не только в исследовательских, но и в учебных целях на базе крупнейших образовательных центров страны.

Синологу, исследующему комбинаторные характеристики языковых единиц, могут помочь корпусы синтаксического типа - the Chinese PropBank[8], аннотированный корпус глаголов китайского языка, и его «корпус-компаньон» - the Chinese Nombank, ориентированные на синтаксический и семантический анализ китайского текста и построение предикатно-аргументных структур.

Развитие компьютерных технологий не только способствовало обновлению старых лингвистических инструментов, но и привело к созданию новых языковедческих ресурсов, к которым

[4] http://app.sinica.edu.tw/cgi-bin/kiwi/mkiwi/kiwi.sh

[5] http://www.livac.org/

[6] www.icl.pku.edu.cn

[7] http://ling.cass.cn/dangdai/corpus.htm

[8] https://www.ldc.upenn.edu/collaborations/current-projects/bolt/annotation/propbank

безусловно относится создание комплексных автоматизированных лексикографических систем, в которых через наблюдение за окружением той или иной лингвистической единицы устанавливаются ее семантические признаки. В настоящее время известно о нескольких реализациях лексических баз данных для китайского языка, крупнейшей их которых является китайский вариант онтологии WordNet.

CHINESE WORDNET (中文詞網)[9] представляет собой компьютерный тезаурус китайского языка, входящий в большую семью аналогичных продуктов WordNet и относящийся к классу лексических онтологий, основным преимуществом которого является свободная доступность для он-лайн поиска. Разработка тезауруса была начата в 2003 году, в настоящее время поддержанием и пополнением CHINESE WORDNET занимается специальная научная группа в Институте лингвистики Тайваньского государственного университета. Для построения китайского варианта WordNet использовались лингвистические ресурсы корпуса Academia Sinica Corpus, однако, как представляется, покрытие лексики и представленные семантические отношения все еще далеко неполны, сведения об объеме тезауруса на сайте разработчиков отсутствуют.

Основными единицами структуры CHINESE WORDNET, на которых задаются семантико-деривационные и семантико-грамматические отношения, как и в других продуктах этой семьи тезаурусов, являются составляющие синсетов (синонимических рядов, кодирующих некоторое понятие) – отдельные лексемы; для китайского языка в ряде случаев установлены немногочисленные отношения синонимии, антонимии, строятся деревья ги-

[9] http://lope.linguistics.ntu.edu.tw/cwn/

понимии (родовидовые отношения) и меронимии (отношения часть-целое), однако примеры употребления лексем синсета в определенном контексте даются далеко не для всех лексических единиц. Учитывая трудности с определением частеречной принадлежности слов в китайском языке, разработчики помещают в одном блоке лексемы различных лексико-грамматических классов, хотя описания, соответствующие разным частям речи, имеют неодинаковую структуру. Прежде всего выделяются отдельные значения слов с толкованиями, для всех полученных лексем определяется базовая парадигма (в широком понимании, т.е. не только морфологические показатели, но и строевая лексика, выполняющая грамматические функции в языковой системе) – производится привязка статей к грамматическому словарю, при этом лексемы различаются не только по частям речи, но и по другим признакам, например, переходности, одушевленности и др., для глаголов дается описание валентностной структуры, в результате синсеты дополняются синонимами и для большинства лексем формируется дерево гипонимии, соответствующее структуре тезауруса.

Этот ресурс значительно расширяет исследовательские возможности анализа дистрибутивных и других свойств слов, особенно китайских глаголов и отглагольных показателей, на основе эмпирических данных, извлекаемых из корпусов текстов. CHINESE WORDNET, являющийся на современном этапе одним из основных инструментов для систем лексикографической обработки естественного языка, также находится в стадии активного пополнения.

Учитывая требование большого объема, как правило, предъявляемое к репрезентативным корпусам текстов, и неоднозначность решения применительно к китайскому языку такой лингвистической задачи, как определение частеречной принад-

лежности слов, автоматическое аннотирование текстовых массивов и разработка алгоритма их полной морфосинтаксической разметки представляются на нынешнем этапе самой актуальной задачей.

3. Основные проблемы китайских корпусов

Появление первых он-лайн корпусов китайского языка сдерживалось долгое время сложностью компьютеризации иероглифической письменности, однако даже после разрешения проблемы ввода иероглифов отсутствие какой-либо разметки в предлагаемых текстовых базах еще долгое время сводило использование корпусов лишь к составлению конкордансов.

Основные принципы организации корпуса китайского языка в целом не расходятся с принятыми в корпусной лингвистике. Специфические проблемы корпусов связаны со структурными особенностями китайского языка как яркого представителя изолирующих языков. В китайском языке, где присутствуют лишь слабые признаки агглютинативной морфологии, слова разных классов не являются четко противопоставленными друг другу на морфологическом уровне, в результате чего учение о частях речи и грамматических категориях не было развито в Китае вообще.

Одной из самых дискуссионных является проблема неоднозначных случаев, когда языковая единица может входить в два или три класса слов, и ее актуальность связана именно с необходимостью морфологической разметки корпуса, где важнейшим фактором оказывается не просто определение набора морфологических признаков в целом, но детальная разработка правил присвоения этих признаков единицам текста.

Определенные трудности вызывает и сегментация китайского текста на отдельные словоформы, связано это с внешней не-

различимостью в китайском языке односложного слова и морфемы в составе сложного слова. Одна и та же единица языка в одном контексте может оказаться словом-членом словосочетания, а в другом контексте функционировать как элемент бинома или полисиллаба, поскольку в китайском словообразовании ведущая роль принадлежит словосложению в чистом виде, без использования каких-либо деривационных элементов. Слова китайского языка демонстрируют плавный переход от слова к словосочетанию с однотипными отношениями между компонентами, при этом степень формальной близости сложных слов к свободным сочетаниям может быть градуированной.

Описанные проблемы даже в условиях постоянно совершенствующихся компьютерных технологий не позволяют добавлять морфологическую, синтаксическую или семантическую информацию о той или иной лексической единице автоматически, требуется ручная разметка корпусов, дальнейшая проверка и внесение синтаксических помет лингвистом, что приводит к трудновыполнимости данной задачи на больших массивах текстов.

Литература

1. *Горелов В.И.* (1989), Теоретическая грамматика китайского языка. М.

2. *Драгунов А.А.* (1952), Исследования по грамматике современного китайского языка. Части речи. М., Л.

3. *Guo Rui* (2002), Studies of parts of speech in Chinese [郭瑞 现代汉语词类研究], Beijing.

4. *Huang Changning* (2002), Corpus Linguistics [黄昌宁 语料库语言学], Beijing.

5. *Shiwen Yu, Hui Wang* (2003), The Semantic Knowledge-base of Contemporary Chinese and it's Applications in WSD, in Proceedings of the second SIGHAN workshop on Chinese language processing, Vol. 17, pp. 112–118. URL:

http://delivery.acm.org/10.1145/1120000/1119266/p112-wang.pdf?ip=212.44.132.21&id=1119266&acc=OPEN&key=4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E6D218144511F3437&CFID=354467140&CFTOKEN=28289888&__acm__=1402434523_41a875a28f1493e084b920984044500a (дата обращения: 03.02.2015)

6. *Yang Xiao-jun* (2006), Survey and Prospect of China's Corpus-Based Research, in Corpus Linguistics Around the World, Language and Computers, Num. 56. Amsterdam, NY, pp.219–232.

References

1. *Gorelov V.I.* (1989), Teoreticheskaja grammatika kitajskogo jazyka. [The Theoretical Grammar of the Chinese Language]. Moscow.

2. *Dragunov A.A.* (1952), Issledovanija po grammatike sovremennogo kitajaskogo jazyka. Chasti rechi. [Studies on the Grammar of the Modern Chinese Language. Parts of Speech] Moscow, Leningrad.

3. *Guo Rui* (2002), Xiandai hanyu cilei yanjiu [Studies of parts of speech in Chinese]. Beijing.

4. *Huang Changning* (2002), Yuliaoku yuyanxue [Corpus Linguistics], Beijing.

5. *Shiwen Yu, Hui Wang* (2003), The Semantic Knowledge-base of Contemporary Chinese and it's Applications in WSD, in Proceedings of the second SIGHAN workshop on Chinese language processing, Vol. 17, pp. 112–118, available at: http://delivery.acm.org/10.1145/1120000/1119266/p112-wang.pdf?ip=212.44.132.21&id=1119266&acc=OPEN&key=4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E6D218144511F3437&CFID=354467140&CFTOKEN=28289888&__acm__=1402434523_41a875a28f1493e084b920984044500a.

6. *Yang Xiao-jun* (2006), Survey and Prospect of China's Corpus-Based Research, in Corpus Linguistics Around the World, Language and Computers, Num. 56. Amsterdam, NY, pp.219–232.

Колпачкова Елена Николаевна
Санкт-Петербургский государственный университет (Россия).
Kolpachkova Elena
Saint-Petersburg State University (Russia).
E-mail: ekolpachkova@gmail.com