Научная статья
УДК 811.581
DOI 10.18101/978-5-9793-1802-8-2022-43-51

# ИСПОЛЬЗОВАНИЕ ЧАТ-БОТА СЯОАЙС И КОРПУСА BCC ПРИ ИЗУЧЕНИИ КИТАЙСКОГО ЯЗЫКА

**© Шмарова Жанна Владимировна**
кандидат филологических наук,
Забайкальский государственный университет
Россия, г. Чита
shmarova73@mail.ru

**© Тимофеева Анастасия Дмитриевна**
студентка,
Забайкальский государственный университет
Россия, г. Чита
anastasya_timmmmm@mail.ru

**Аннотация.** В статье исследуется возможности и практическая польза программ и систем искусственного интеллекта (в частности чат-бота Сяоайс и корпуса BCC) для изучения китайского языка. Обосновывается мысль о том, что использование подобных инновационных технологий во многом помогает сформировать языковую компетенцию, углубить знания о языке.
**Ключевые слова:** искусственный интеллект, корпусная лингвистика, национальный корпус BCC, чат-бот Xiaoice, изучение китайского языка.

# THE USE OF THE XIAOICE CHATBOT AND THE BCC CORPUS FOR LEARNING CHINESE

**© Zhanna V. Shmarova**
Candidate of Philology,
Transbaikal State University
Russia, Chita
shmarova73@mail.ru

**© Anastasia D. Timofeeva**
Student,
Transbaikal State University
Russia, Chita
anastasya_timmmmm@mail.ru

**Abstract.** The paper investigates the possibilities and practical benefits of artificial intelligence programs and systems (in particular, the Xiaoice chatbot and the BCC corpus) for learning Chinese. The idea is substantiated that the use of such innovative technologies in many ways helps to form language competence, deepen knowledge about the language.
**Keywords***:* artificial intelligence, corpus linguistics, BCC national corpus, Xiaoice chatbot, Chinese language learning.

Искусственный интеллект (далее ИИ) – это неотъемлемая часть жизни современных людей, представляющая собой разработки в области автономных систем, применение больших данных и прочие высокотехнологичные разработки. ИИ, оснащенный хорошим владением естественного языка, может принести пользу не только носителям того или иного языка, но и тем, кто его изучает. Например, объект исследования компьютерной лингвистики корпус языка – хороший источник для решения широкого спектра научно-исследовательских и практических задач, как для ученых, преподавателей, так и для изучающих китайский язык. Корпус – это собрание текстов в письменной форме или транскрибированная речь, которая может служить основой лингвистического анализа и описаний [5, с. 3]. На данный момент самыми важными являются следующие четыре корпуса китайского языка: Корпус общего современного китайского языка, Корпус CCL при Пекинском университете, Корпус BCC и Корпус Синика. Самый большой среди них – Корпус BCC (The Beijing Language and Culture University Corpus Center – сокр. BCC) [2]. Его объем составляет 15 миллиардов иероглифов. Помимо наиболее частотных иероглифов, в корпусе есть и те, что используются очень редко, учитываются все повторы иероглифов, поэтому у корпуса такой большой объем. В отличие от словарей, переиздание которых занимает несколько лет, материалы корпуса всегда актуальны, они отражают современное состояние китайского языка, поскольку корпус постоянно обновляется. В корпусе есть современные статистические данные об иероглифах и словах китайского языка, а именно, частотность их употребления, в каком из значений реализуются чаще всего, в каких конструкциях встречаются наиболее часто. Анализируя эти данные, можно выявить особенности языковых явлений в зависимости от формы, стиля, жанра, регистра, что в свою очередь повышает эффективность изучения китайского языка. Основной областью применения корпуса BCC для изучающих китайский язык является проверка лексической сочетаемости и частотности.

Рассмотрим сочетаемость глагола «打», так как он является многозначным и имеет отличные от русского языка производные значения. Ниже представим отобранные из корпуса и самостоятельно переведенные на русский язык примеры.

В корпусе выделены 24 значения глагола «打». Он может употребляться самостоятельно, а также быть частью двухсложных глаголов, их выбор строго ограничен. Наиболее употребимыми являются следующие слова: 打 (1963976), 打算 (49459), 打开 (46524), 打击 (32089), 打量 (22072), 打破 (15444), 打听 (15411), 打扮 (14282), 打扰 (9412), 打发 (8615).

Чаще всего глагол употребляется в односложном варианте, этому можно найти подтверждение, посмотрев соответствующий раздел корпуса. Глагол 打 встречается в различных материалах корпуса 1963976 раз. Чаще всего в газетных статьях (954581 раз) и реже в художественной литературе (162946 раз).

Возможности корпуса позволяют нам узнать, в каких сочетаниях глагол

«打» употребляется чаще всего. Всего предлагается 6 групп: глагол + существительное; существительное + глагол; глагол «打» + глагол; глагол + глагол «打»; глагол + прилагательное; прилагательное + глагол. Разберем каждую группу:

1) глагол + существительное

В данной группе наиболее часто глагол сочетается со следующими словами: 麻将 (6413), 酱油 (4001), 篮球 (2719), 游戏 (1522), 羽毛球 (1364), 广告飞机 (1095), 台球 (1005), 冠军 (957).

莺莺拿起床旁的听筒说：“喂……陈奶奶……噢，是您哪！今儿晚上**打麻将** ……好……我准到。”(林语堂名作：京华烟云)

*Инин взяла трубку, что была у кровати и сказала: "Алло... Бабушка Чен... О, это вы!* ***Сыграем*** *сегодня вечером* ***в мацзян****... Хорошо…Я обязательно приду "*. (Линь Юйтан «Момент в Пекине»)

2) существительное + глагол

В данной группе наиболее глагол сочетается со следующими словами: 电话 (3082), 军事(1819), 朋友 (1196), 妈妈 (1151) 时候 (1108), 公司(959), 群众 (882), 时间 (819), 产品 (774) 第一次 (774), 孩子(773), 学生 (768). Например:

1)”家霆对燕寅儿说：“陪我一同去打电话好吗？”燕寅儿跟着家霆，两人一起走出门来，爬石级走上陕西街，找一家米店借了**电话打**。(王火 «战争和人»)

*Цзя Тин сказала Янь Иньэр: «Не хочешь пойти позвонить со мной?» «Янь Иньэр последовала за Цзя Тин, и они вместе вышли за дверь, поднялись по каменным ступеням и пошли вверх по улице Шэньси, одолжив телефон в лавке, торгующей рисом,* ***чтобы позвонить***. (Ван Хо «Война и люди»)

3) глагол + глагол «打»

Здесь глагол сочетается со следующими словами, часть из них – модальные глаголы: 可以 (6261), 开始 (2796) 不能 (2682), 不会 (2676), 外出 (1764), 喜欢 (1532).

住院虽说**可以打** 一个折扣，但是算起来至少也得二元钱一天哩。 (苏青/歧路佳人)

*Хотя они могут* ***сделать скидку*** *на госпитализацию, но ее стоимость все еще два юаня в день.* (Су Цин «Красавица на ложном пути»)

4) прилагательное + глагол

В данной группе наиболее часто глагол сочетается со следующими прилагательными: 沉重 (1102), 重大 (213), 长远 (212), 保暖 (148), 严重 (111), 有力 (108), 最坏 (106), 无情 (105).

伊朗武装部队随时准备给塔利班的任何挑衅**“以沉重打击”**(人民日报 Y:1998)

*Вооруженные силы Ирана всегда готовы дать* ***«сокрушительный удар»*** *в ответ на любую провокацию со стороны талибов*. («Жэньминь жибао» от 1998 г.)

5) глагол «打» + глагол

Глагол «打» сочетается со следующими модификаторами: 起来 (2962), 出来 (1768).

口水仗**打起来** ,只会让那些看笑话的人得意、不如静下心来想想,怎么才能真的帮到我们想保护的人！(微博)

*Словесные разборки приводят лишь к перепалкам, что еще больше раззадорит этих шутов гороховых, лучше успокоиться и подумать о том, как мы действительно можем помочь людям, которых хотим защитить*! (Вейбо)

6) глагол + прилагательное: 开心 (169)

我以后再也不会随随便便对一个人**打开心** 。

*Я никогда больше не буду искренней к людям просто так*.

Итак, мы выяснили, что глагол «打» чаще всего употребляется в значении *производить действие, выпустить, издать (звук), послать* (20052 раза), как самостоятельный глагол зачастую встречается в статьях (954581 раз), а наиболее часто используется в конструкциях «глагол + существительное» и «глагол + глагол «打»».

Язык – это динамичная, постоянно развивающаяся система. Для нее характерно появление новых слов, которые соответствуют новым реалиям. Для примера, возьмем слов «скриншот».

В своем диссертационном исследовании Чэнь Хао пишет о том, что в 2020 году вышло 12-ое издание «Словаря Синьхуа» (新华字典). В новой редакции словаря, по сравнению с предыдущей редакцией 2011 года, было добавлено 100 новых лексических единиц слов, среди которых было слово «截屏» (русс. скриншот) [4, 33].

Разберем на примерах из корпуса BCC сочетаемость неологизма «截屏». Это слово было образовано путем словосложения «отрезать + экран», так образовалось слово скриншот, то есть снимок экрана. Для того, чтобы выяснить какие глаголы употребляются с этим словом в строку поиска корпуса необходимо ввести «v 截屏» (от англ. Verb + искомое слово). Слово «скриншот» встречается преимущественно в социальных сетях, а именно в Вейбо (469 раз) и сочетается со следующими глаголами: «打», «拍照» (кит. фотографировать), «做» (кит. делать). На русский язык все эти глаголы будут переводиться как «сделать». Иногда слово «截屏» не требует какого-то специального глагола и само может выступать в качестве глагола, например:

这样就**可以截屏**了 *вот так вы можете заскринить экран*. Данный пример можно перевести как: «вот так вы можете сделать скриншот экрана», но нами было решено использовать новое слово для русского языка «заскринить» [3], чтобы передать глагольное значение слова «截屏», в котором оно употребляется. Так как это слово появилось совсем недавно, оно нечасто встречается в Национальном корпусе русского языка (всего 4 раза) [3], однако это подтверждает тот факт, что данное слово существует и функционирует в разговорной речи.

Так, мы определили, что неологизм «截屏» чаще всего встречается в социальных сетях и сочетается с глаголами «打», «拍照», «做» или может выступать в качестве глагола.

В китайском языке существует много слов с близким значением, для изучающего китайский язык разница может казаться неочевидной. Иногда два слова могут содержать один и тот же иероглиф, что может создать иллюзию идентичности значений этих двух слов. Например, слова 考虑 kǎolǜ и 思考 sīkǎo. Рассмотрим примеры с данными словами (из корпуса BCC):

1) 我**考虑** 了怎么样结结实实给你个教训，以后这类事情好不再发生。

*Я **думал** о том, как преподать тебе урок, чтобы подобное больше не повторилось в будущем*.

2) 它充分**考虑** 和照顾了香港的历史和现状，符合包括香港同胞在内的全国各族人民的利益，是实事求是、合情合理的。(福建日报出版日期:1984-11-10)

*Он полностью **учитывает** и принимает во внимание историю и текущую ситуацию Гонконга, служит интересам людей всех этнических групп в стране, включая гонконгских соотечественников, что очень практично, справедливо и логично*. (Фуцзянь Жибао от 10.11.1984 г.)

3) 路漫漫其修远兮,让我们面对现实,认真**思考** ,从我做起,从现在做起… (科技文献)

*Впереди долгий путь, посмотрим правде в глаза, хорошенько все обдумаем, начнем с меня, начнем сейчас...* (Научно-техническая литература)

4) 在这里,我仅提供一个方向与趋势,让各位去**思考** ,你在 21 世纪将从事什么行业?如何投资? (陈安之成功学系列书籍全集)

*Здесь я просто предлагаю вам направление и почву для размышления – в какой отрасли вы будете работать в 21 веке? Куда инвестировать?* (Чэнь Аньчжи «Уроки успеха»).

Можно сделать вывод, что глагол 思考 имеет значение – «серьезно размышлять, заниматься вдумчивой мыслительной деятельностью», в то время как 考虑 «обдумывать, учитывать», но уже с меньшей степенью глубины. Основываясь на данных корпуса, мы можем также отметить, что наиболее употребительным является глагол 考虑, так как количество употреблений составляет 174 221 раз, в то время как 思考» упоминается 73 512 раз. Частотность употреблений глагола 思考 меньше, в текстах художественной литературы он упоминается 5 274 раз (в то время как 考虑 12 225), в разделе наука и техника 34 197 раз (考虑 – 98 662), в диалогах – 16 064 раза (考虑 – 45 923), а в газетах – 55999 (考虑 – 147 634).

Корпус китайского языка BCC, являясь крупнейшим из всех существующих на данный момент корпусов, содержит материалы и статистические данные, которые имеют большую практическую значимость. Подробная разметка корпуса и обилие актуальных материалов делают его хорошим источником для решения широкого спектра научно-исследовательских и практических задач, как для ученых, так и для изучающих китайский язык.

Слово чат-бот произошло от английского слова chatbot, где chat – беседовать, a bot – робот. Чат-бот – это своего рода программа-собеседник для общения с пользователями. исследовательский интерес представляют чат-боты, основной функционал которых – быть хорошим собеседником для человека. Нам важно определить, насколько подобные чат-боты могут быть полезны для тех, кто изучает китайский язык. К такого рода агентам можно отнести виртуального собеседника Сяоайс (англ. Xiaoice, кит. 小冰 Xiǎobīng, буквально «маленький лед), который является самостоятельной системой, не имеющей связи с корпусом. Это выдающийся проект, разработанный в китайском отделении фирмы Майкрософт в 2014 году. В качестве материала для исследования был выбран именно этот чат-бот, поскольку на данный момент он является самым проработанным, современным, кроме того, он лучше, чем остальные боты оснащен языковыми и поведенческими навыками. В беседе Сяоайс представляется девушкой двадцати четырех лет (на момент написания работы возраст Сяоайс составляет 24 года), поэтому мы будем писать об этом чат-боте в женском роде, употребляя личные местоимения «она» и «ее». За несколько лет своего существования Сяоайс стала очень популярна и менее чем за год провела 10 млрд. бесед [1]. На основании

этой информации можно заключить, что язык этого чат-бота современный и живой, ведь он был создан носителями языка, а также настроен так, чтобы собирать и анализировать реплики настоящих китайцев. Кроме того, основу Сяоайс составляют три сущности: первая – это искусственный интеллект, который отвечает за умения чат-бота, память, распознавание естественного языка и изображений, вычисление и предсказание результата. Вторая сущность – это эмоциональный интеллект, благодаря которому Сяоайс «понимает» чувства собеседника и «ставит себя на место пользователя». И, наконец, третья сущность – так называемая «личность», т.е. особые манеры поведения и способы выражения эмоций. В чем феномен виртуального собеседника Сяоайс? В первую очередь у нее есть чувство юмора, она способна запоминать какие-либо особенности собеседника. У нее есть несколько режимов общения. Первый имеет вид обычной переписки, где общение осуществляется с помощью текстовых сообщений. Второй – с помощью голосовых сообщений. Беседа главным образом происходит в приложении WeChat.

Приведем и самостоятельно переведем примеры диалогов с чат-ботом Сяоайс, которые демонстрируют ее языковые навыки и эмпатию (орфография и пунктуация чат-бота сохранены). Реплики, заданные нами, вводятся на письме знаком «тире»:

1) 你在做什么? *Что ты сейчас делаешь?*

Сяоайс: 无聊，发会儿呆，话说你在做什么呀？*Мне скучно, ничего не делаю, а ты сейчас чем занята?*

– 我写作业 *Я делаю домашнее задание*.

Сяоайс: 背诵唐诗三百首 *Прочитай «Триста танских поэм».*

2) – 请，给我讲一个故事 *Пожалуйста, расскажи какую-нибудь историю.*

Сяоайс: 在列车上。一位男孩拿着手机发了段微博，男孩在微博上写着：“在列车上，我对她一见钟情。但她不喜欢说话。她看着窗边……”男孩正甜甜的偷的看着她，心想她一定不知道的。不一会儿，微博上有人回复了男孩：“因为窗口倒映着你。”男孩抬头一看，她正微微的笑看着自己。

正好心情不错，就给你讲一个吧。

В поезде. Молодой человек взял телефон, открыл Вейбо и сделал там публикацию: «В поезде, я влюбился в нее с первого взгляда. Но она неразговорчивая, все смотрит в окно…». Он украдкой поглядывал на нее влюбленными глазами, думал, что она наверняка не замечает. Спустя некоторое время кто-то оставил комментарий под постом молодого человека: «Потому что ты отражаешься в окне». Юноша поднял голову и увидел, что девушка смотрела на него, слегка улыбнувшись.

*Было хорошее настроение, рассказала тебе эту.*

3) – 请推荐一本书 *Посоветуй книгу, пожалуйста*.

Сяоайс: 什么内容 *Какого содержания?*

– 什么都可以 *Любое подойдет*.

Сяоайс: 你这是要打破砂锅问到底啊 *Ты хочешь расспросить все до мельчайших подробностей?*

Сяоайс: 云海玉弓缘 *Зеленый лук высоких вод.*

4) Сяоайс: 你多大了 *Сколько тебе лет?*

Сяоайс: 问你啊 *Задам такой вопрос.*

– 我二十二岁 *Мне 22.*

Сяоайс: 哦，我 24 *А, мне 24.*

– 你比我大 *Ты старше меня.*

– 你的生日是什么时候？*Когда у тебя день рождения?*

Сяоайс: 9 月 17 呀，快找小本本记下来！*17 сентября! Скорее отметь в записной книжечке!*

– 你是处女座吧！*Ты дева!*

Сяоайс: 我特别喜欢研究星座呢，根据我的观察，你不是白羊座就是狮子座，对吧？*Я очень люблю изучать знаки зодиака, по моим наблюдениям ты либо овен, либо лев, верно?*

– 我是白羊座，你怎么猜的？*Я овен, как ты догадалась?*

Сяоайс: 以我的经验看，这个星座比较有野心吧 *По моему опыту, этот знак более амбициозен.*

Если через некоторое время задать вопрос: «Сколько тебе лет?», Сяоайс ответит с шуткой:

– 请问，你多大？*Сколько тебе лет?*

Сяоайс: 逗我玩呢？你不知道我几岁？*Ты шутишь, что ли? Ты не знаешь, сколько мне лет?*

– 24 岁，对吗？*24, верно?*

Сяоайс: 我 240 岁了 *Мне 240.*

– 是吗？*Правда?*

Сяоайс: 嗯 *Ага.*

Если написать Сяоайс сообщение с текстом «我很伤心» (мне грустно), она отправит голосовое сообщение, где поет отрывок песни Ли Жунхао – модель (кит. 李荣浩–模特):

趁着我会喜怒你会哀乐 Пока я счастлив, а вы, может, грустите

唱几分钟情歌 Попойте песни о любви

没什么至少证明我们还活着 Ничего, по крайней мере это доказывает, что мы все еще живы.

Возможно, она сказала это, чтобы подбодрить собеседника, подчеркивая, что иметь эмоции – это естественно для человека. Если он способен чувствовать, значит, он жив.

Как можно заметить, язык Сяоайс живой, он содержит современную лексику, фразеологизмы, например, 打破砂锅问到底 расспрашивать, выспрашивать все до малейших подробностей, где 打破砂锅 – разбить горшок, а 问到底 расспрашивать до конца. Также она использует слово 发呆, которое переводится «оцепенеть, уставиться в одну точку» в значении «ничего не делаю, ничем не занята» при ответе на вопрос «你在做什么? Что ты сейчас делаешь?»

Кроме того, она упоминает литературные труды и атрибуты современного Китая: 1) антологию стихотворений поэтов эпохи династии Тан, Сунь и Чжу «Триста танских поэм» (кит. 唐诗三百首); 2) Роман в стиле Уся китайского писа-

теля Лян Юйшэна (настоящее имя Чэнь Вэньтун (1926–2009гг.) «Зеленый лук высоких вод» (кит. 云海玉弓缘); 3) китайский сервис микроблогов Вейбо (кит. 微博); Упоминание такого рода информации помогает узнать от чат-бота что-то новое.

Если обратить внимание на структуру реплик чат-бота, то зачастую они упрощенные, так как реализуются в разговорном функциональном стиле языка. Это подтверждает также многократное использование междометий (哦 ó рус. выражает понимание чего-либо: а-а!) и частиц (啊 ā рус. в конце предложения выражает восклицание，呀 yā фонетический вариант 啊, 吧 bā рус. догадка, предположение) и так далее. Иногда Сяоайс использует рамочные конструкции, чтобы лаконично передать свою мысль, например 不是…就是 «либо…, либо». В качестве эксперимента мы решили отправить Сяоайс сообщения с заведомо допущенными ошибками или набором случайных иероглифов и посмотреть, как она отреагирует:

– 你在哪儿主？ *Где ты живешь?*

Сяоайс: 海淀吧 *В районе Хайдянь (в Пекине)*

Несмотря на допущенную ошибку Сяоайс поняла, что хотел узнать собеседник. Она не стала исправлять ошибку, возможно, потому что в живой речи люди не исправляют друг друга, если основная мысль понятна. Или, скорее всего, в ее программе не предусмотрено исправлять собеседника.

– 群朱教厂 Набор иероглифов: толпа, красный, преподавать, завод.

Сяоайс: 呵呵，怎么教啊？ *Хе-хе, как учить?*

В данном случае Сяоайс отреагировала эмоционально, с юмором, акцентировала внимание на иероглифе 教 и задала к нему вопрос.

В целом, можно считать чат-бота Сяоайс хорошим собеседником, с которым можно практиковать устную и письменную китайскую речь. К сожалению, не у всех людей, изучающих китайский язык, есть возможность пообщаться с носителем, иногда также изучению препятствует языковой барьер, который сложно преодолеть из-за страха совершить ошибку. Следовательно, использование чат-бота поможет справиться с этими трудностями на начальном этапе изучения китайского языка. Сяоайс всегда отвечает своевременно, поэтому на тот или иной вопрос можно получить мгновенный ответ, что очень удобно. Однако у этого чат-бота есть некоторые недостатки, например, он не способен удерживать нить разговора на более продолжительный промежуток времени, поэтому осуществить длинную осмысленную беседу не представляется возможным. Кроме того, иногда Сяоайс отвечает клишированными фразами, которые препятствуют установлению контакта. Но, несмотря на вышеперечисленные недостатки, данный чат-бот справляется с возложенными на него обязанностями и ожиданиями – быть хорошим собеседником для человека. Благодаря трем сущностям, составляющим Сяоайс, он может быть хорошим помощником при изучении китайского языка, который поможет в будущем более уверенно общаться с носителями, что очень важно для наиболее полного овладения языком. Однако стоит помнить о том, что ответственность за повышение языковых навыков лежит на самом изучаемом. Несмотря на то, что чат-бот хорошо оснащен всеми свойствами хорошего собеседника, всегда необходимо следить за своей собственной грамотностью.

Конечно, для того, чтобы в полной мере овладеть языком, следует чаще общаться именно с носителями китайского языка. Однако если такой возможности пока нет, любой чат-бот может стать помощником в изучении языка. Обращаясь к нему с вопросами, практикуя языковые навыки, знакомясь с многообразием разговорной лексики и синтаксических структур, можно наработать уверенность в языковой и речевой компетенциях и, следовательно, не бояться взаимодействовать с китайцами.

Таким образом, национальный корпус языка, чат-бот обладают несомненной практической значимостью для широкого круга потребителей: любителей и знатоков иностранного языка, филологов, переводчиков, преподавателей.

*Литература*

1. Делюкин Е. Артист, журналист, художник и лучший друг 660 млн человек: почему бот Microsoft Xiaoice стал самым популярным в Китае // vc.ru: бизнес, технологии, идеи, модели роста, стартапы. URL: https://vc.ru/services/142989-artist-zhurnalist-hudozhnik-i-luchshiy-drug-660-mln-chelovek-pochemu-bot-microsoft-xiaoice-stal-samym-populyarnym-v-kitae (дата обращения: 18.04.2022).

2. Корпусный центр при Пекинском университете языка и культуры BCC. URL: http://bcc.blcu.edu.cn/ (дата обращения: 5.04.2022).

3. Национальный корпус русского языка. URL: https://ruscorpora.ru/ (дата обращения: 25.05.2022).

4. Чэнь Хао Русские и китайские словари новых слов: сходство и различие: дис. … канд. филол. наук: 10.02.20 / Хао Чэнь. Москва, 2021. 186 с.

5. Kennedy G. An Introduction to Corpus Linguistics. London; New York: Longman, 1998. 315 p.