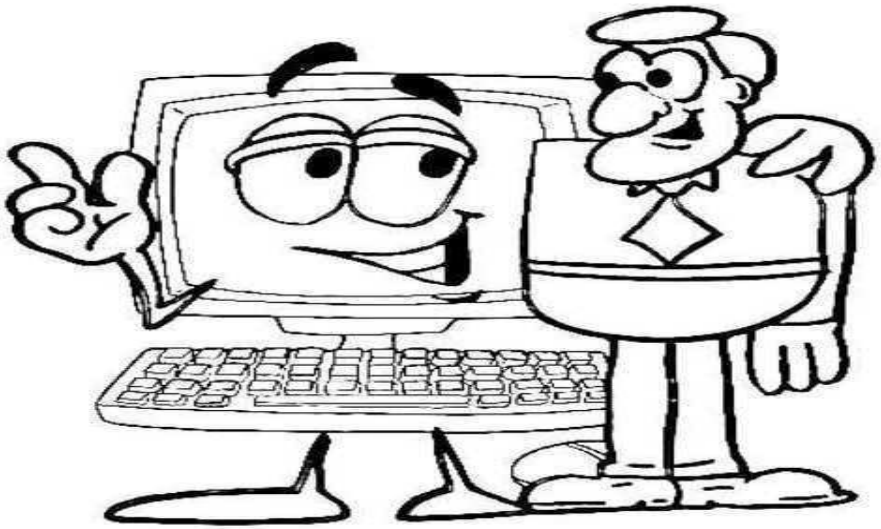
Гибкий Павел

Корпусы текстов китайского языка

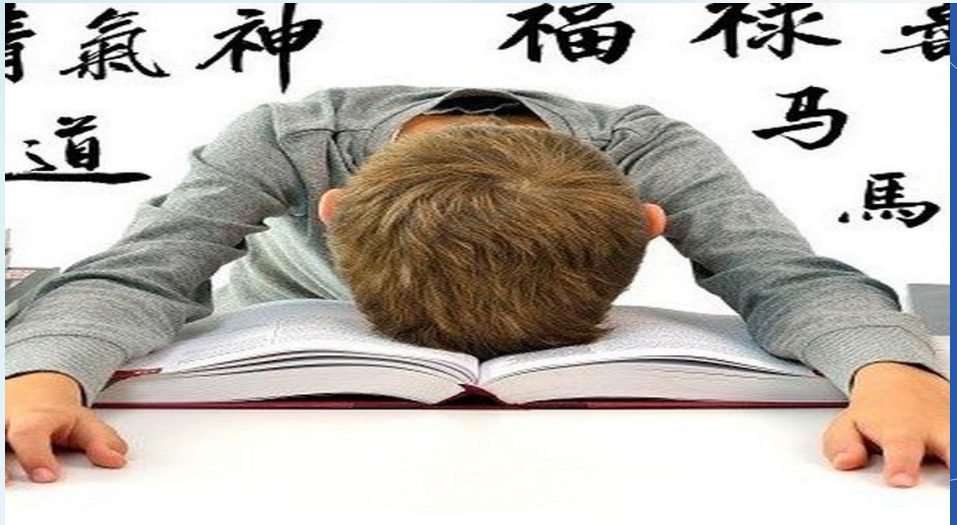


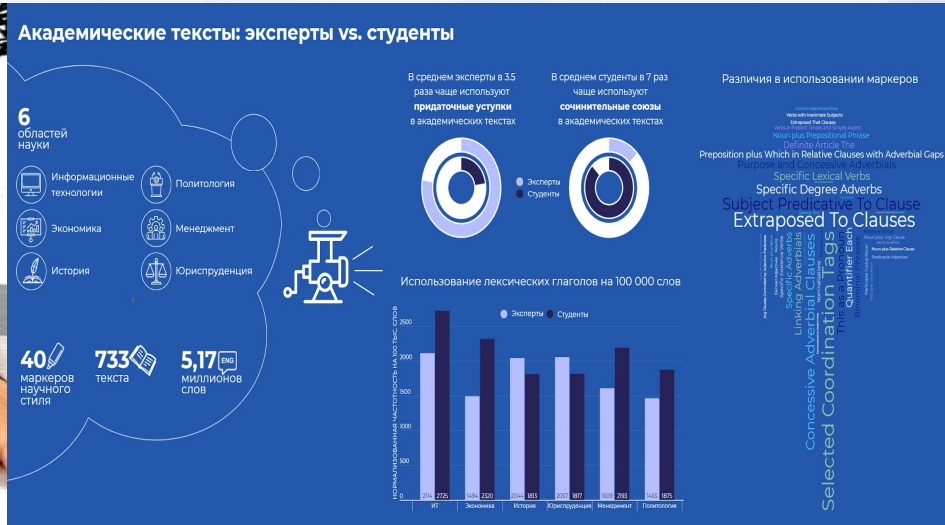
«Конец 1980 – середина 1990-х гг., создаются **корпусы текстов на национальных языках** в разных странах, в том числе и в Китае» [1].

人民日报

Предвестник корпусов китайского языка

Собрание китайских текстов для исследования частотности 语体文应用字汇 («Сборник текстов для изучения единиц разговорного стиля языка») [3].





Первый китайский лингвистический корпус

«人民日报»标注语料库（**корпус газеты** «Жэньминьжибао»(**1999** г. [3]).

Современные корпусы китайского языка

- **«Лингвистический корпус китайского языка Пекинского университета языка и культуры** (BCC).
- **Center for Chinese Linguistics** (CCL) .
- **Chinese Corpus online (语料库在线,** языковые материалы с 1919 года**)» [2].**
- **НКРЯ (Русско-китайский параллельный корпус Национального корпуса русского**

BCC

Крупнейший корпус китайского языка в мире (15 млрд иероглифов).

Chinese Corpus online

Программы **автоматической сегментации текстов**, **частеречной** разметки слов, **подсчета частотности слов** и **разметки** пиньиня **(100 млн**

4

CCL

Корпус **современного, древнекитайского языков, китайско-английский корпус (500 млн иероглифов).**

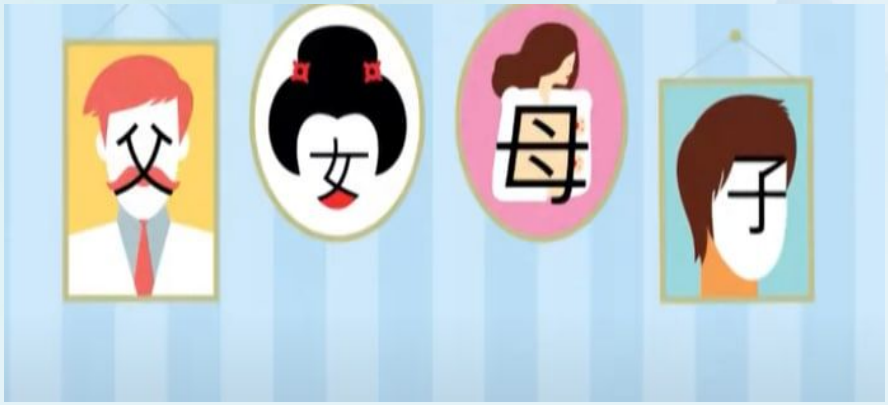
3

НКРЯ

- **самый большой** открытый параллельный корпус русского и китайского языков.

Особенности китайских корпусов текстов

«**Не все** иероглифы китайского языка характеризуются **высокой встречаемостью** в текстах. Характерной особенностью иероглифов является их несоответствие буквенно-словесным универсалиям» [5].

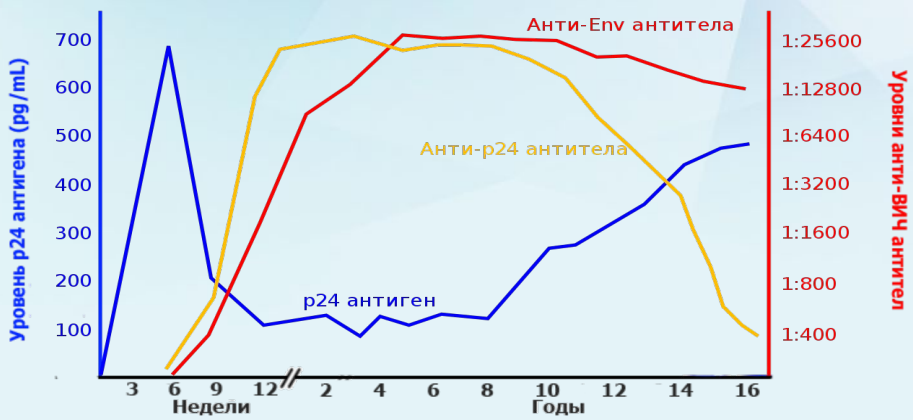


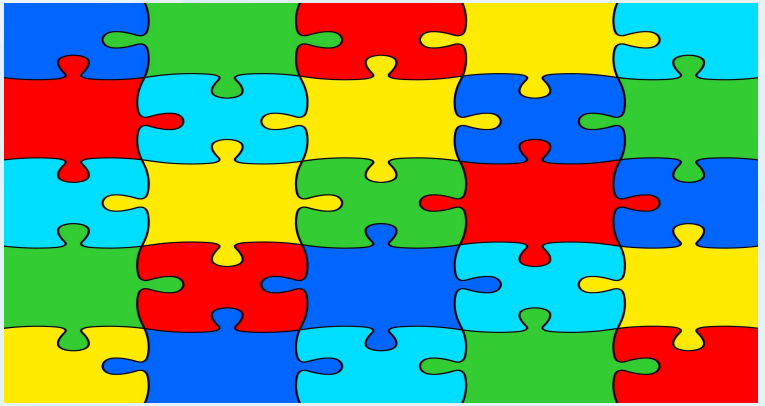
Проблемы ранних корпусов:

1. Большая часть данных в**водилась вручную**, небольшие размеры корпусов.

Проблемы ранних корпусов:

2. В силу использования **разными корпусами разных методов** автоматической сегментации получались **разные результаты**.



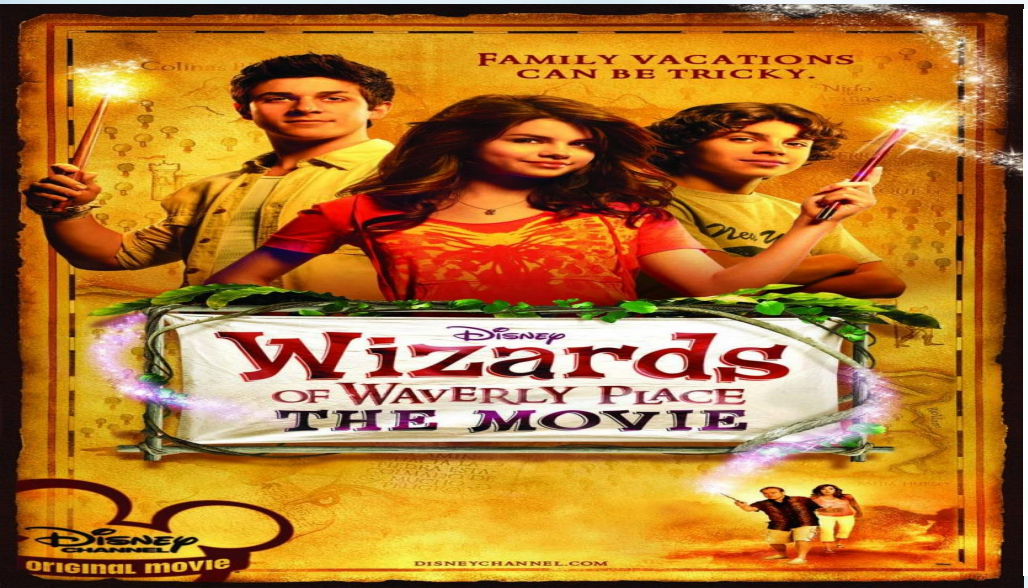


Решение проблем:

1) **Разработка национального стандарта** GB -13715 «Современная сегментация китайского слова ...» .

2) **Составление первого масштабного корпуса китайского языка** — 语料库在线 [4] (1991 год).

3) **Составление** CLL (783 463 175 знаков).



Список литературы

1. **Баркович, А. А. Лингвистические корпусы китайского языка: функциональный аспект / А. А. Баркович, Ван Цин // Вестник МГЛУ. – 2015.. – . – Т. № 5, № (78). – С. 105 – 113.**
2. Фэн, Юэ. Специфика корпусных исследований в современном китайском языкознании / Юэ Фэн, Ван Цин // Вестник МГЛУ. – 2020. – . – Т. 3, № 832. – С. 159 – 172.
3. 陈鹤琴 语体文应用字汇 = Сборник текстов для изучения языковых единиц разговорного стиля [Электронный ресурс]. – Режим доступа : http://book.ln.chaoxing.com/ebook/read_11378 972.html. – Дата доступа : 25.04.2015.
4. 人民日报»标注语料库 = Размеченный корпус газеты «Жэньминьжибао» [Электронный ресурс]. – Режим доступа : http://ling.cass.cn/yingyong/courses/ corpusbase.htm. – Дата доступа : 25.04.2015.
5. 字、词 – 现代汉语 = Соотношения между символами, словами и морфемами [Электронный ресурс]. – Режим доступа : http://www.yyxx.sdu.edu.cn/chinese/wt/main04-03.htm. – Дата доступа : 25.04.2015.