Русско-китайский параллельный корпус НКРЯ



Что такое параллельный корпус?

• Корпус:

- Собрание текстов
- Каждое слово в котором размечено (т.е. есть грамм. и другая информация)
- По каждому слову можно вести поиск
- Параллельный:
 - Тексты собраны на двух языках и выровнены по предложениям

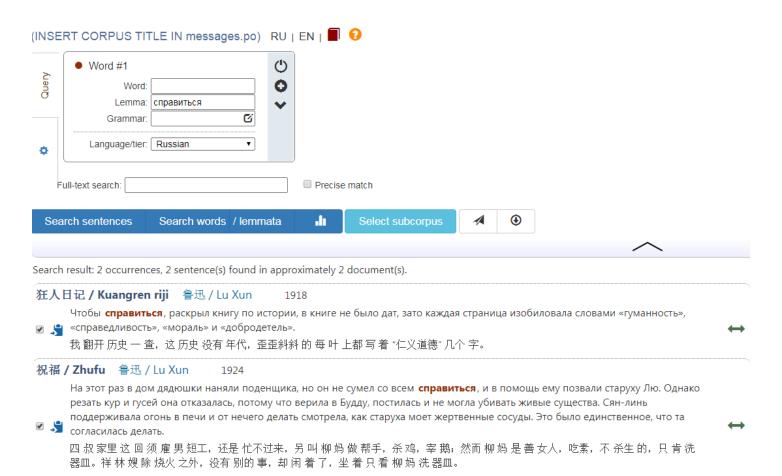
Cui bono?

- Изучающим иностранный язык (и преподавателям)
- Переводчикам
- Компьютерным лингвистам и программистам

Наш корпус

- Где:
 - https://linghub.ru/rnc parallel chinese/search
 - http://www.ruscorpora.ru/new/search-para-zh.html (хромает)
- Кто:
 - Институт русского языка РАН
 - ВШЭ (ФиКЛ, ФКН и ДИЯ)
 - МГУ
 - РГГУ (Институт Конфуция)
 - МФТИ/Huawei
 - СПбГУ
 - Чжэцзянский университет
 - Хэйлунцзянский университет

Как это выглядит сейчас?

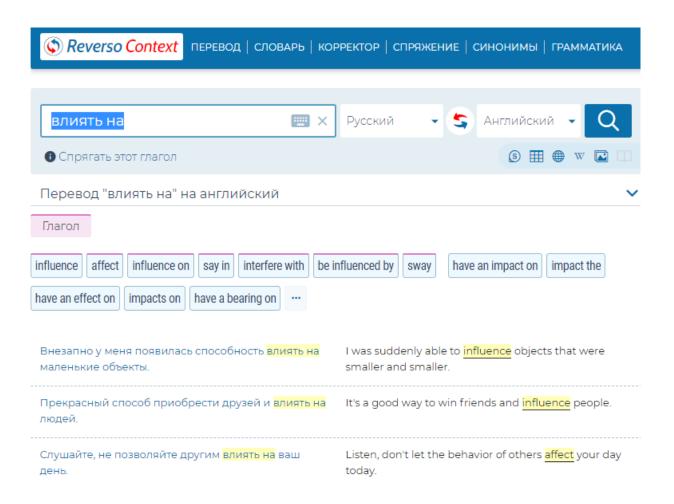


- Языки:
 - Русский
 - Китайский (иероглифы)
 - Китайский (пиньинь)
- Поиск по:
 - Словоформе
 - Начальной форме (lemma)
 - Грамм.показателю
- Можно выбрать подкорпус

Размер корпуса

- 177 000 русских слов
 - Скоро будет + 450 000 слов
- Произведения XIX нач. XX вв.
- 7 русских произведений, 3 китайских

Наша цель:



- Большой
- Разнообразный
- Удобный
- Популярный

Наши задачи (для китаистов)

- Выравнивать тексты (готовить их к формату корпуса)
- Искать новые ПАРАЛЛЕЛЬНЫЕ тексты
- Проверять качество алгоритмов:
 - Словоделения
 - Пиньиня
 - Выравнивания
- Поддерживать корреспонденцию с китайскими коллегами
- (в будущем) заниматься СММ

Наши задачи (для программистов)

- администрирование сайта на сервере
- фронт-энд: создать две новые версии сайта:
 - более простой интерфейс на русском языке для пользователейнелингвистов + прикрутить нейронки
 - интерфейс на китайском языке
- адаптация других корпусов под наш формат
- встроить автоматическую систему присвоения транскрипции иероглифам с хорошей точностью
- делать любые другие исследовательские и технические работы с нашим корпусом. Инициатива приветствуется!

Формы кооптации

- Практика
- Волонтерство
- Ярмарка проектов
- НУГ (научно-учебная группа)
- Оплачиваемая «практика»
- Кооптация со стороны (классический аутсорс)

Спасибо за внимание!

Кирилл Семенов, ФиКЛ:

kir.semenow@yandex.ru

Vk: cyrilsemenov

Tg: @cyr_sem