СИНОЛОГИЯ

С.В. Зинин

Торонто

Новый интерактивный сетевой конкорданс *Чуньцю* и *Цзочжуани*

Проект ZZSTATS [6] представляет собой интерактивный сетевой конкорданс текстов Чуньцю и Цзочжуань (далее ЧЦ и ЦЧ). Это третий по счёту сетевой конкорданс для ЦЧ и второй для ЧЦ. Он отличается от двух предыдущих[1] как структурно, так и содержательно[2], и поэтому представляет собой самостоятельную ценность. Новый конкорданс будет незаменим для многих видов текстологических исследований. Хотя в настоящее время проект представляет собой в основном конкорданс, он развивается как интерактивная среда для анализа нарративов. Настоящая работа представляет собой краткое описание проекта.

В качестве цифровой версии текста ЧЦ и ЦЧ в ZZSTATS была избрана версия китайской части сетевой энциклопедии Wikimedia в кодировке UTF-8 (см. [4]). Сравнительный анализ показывает, что для сетевых конкордансов китайских классических текстов огромное значение имеет выбор версии текста и даже его кодировки. Версия текста во многом определяет результаты поиска по тексту, а также его статические характеристики: так, некоторые элементы текста могут присутствовать в одних версиях и отсутствовать в других[3], количество найденных фрагментов для одного и того же знака также может отличаться, даже если варианты текста по длине будут практически одинаковы. Некоторые иероглифы в различных изданиях текста могут присутствовать в своих вариантах. Конвертация некоторых особо редких знаков, часто встречающихся в древних текстах, может происходить с ошибками (знак заменяется на код UTF-8, выставляющийся «по умолчанию»), и это тоже сказывается на результатах поиска.

В то же время, наличие нескольких сетевых конкордансов для одного и того же памятника позволяет провести сравнительный анализ в отношении отдельных иероглифов. В итоге оказывается, что очень часто привычное филологическое утверждение типа «в памятнике X иероглиф

Y употребляется Z раз» должно обязательно сопровождаться пояснением – «в такой-то версии памятника». Это утверждение может показаться банальным. Однако, в период, когда печатные конкордансы существовали в единственном варианте, или же исследователь опирался на избранную им версию текста памятника, всё ограничивалось ссылкой на издание соответствующего индекса в библиографии. Сравнительный текстологический анализ производился только в исключительных случаях и был затруднён объёмом материала. Эта проблема начинает исчезать, по мере того, как всё большее количество текстов становится доступным в своих цифровых вариантах.

Вряд ли следует ожидать появления единственного «наиболее авторитетного» цифрового варианта известных памятников. Появление же конкордансов различных версий памятников заставляет исследователей учитывать все различия в диапазоне всего текста, а не в отдельных фрагментах, как это происходило ранее. В случае ZZSTATS, главное назначение которого – нарративный анализ – наличие или отсутствие того или иного редкого знака не является столь важным, как для потенциального критического издания текста.

В настоящее время в проекте выделяются четыре основные функциональные группы (мы приводим также их русские и английские названия, так как интерфейс проекта выполнен на английском языке):

1) Собственно тексты памятников и конкорданс (**Texts**);
2) Поиск по иероглифам и их чтениям (**Search**);
3) «Корзина» иероглифов (**Bag**);
4) Общие статистические данные (**Stats & Freqs**).

Texts

В разделе «Тексты» пользователю предоставляется возможность доступа к собственно тексту памятников, в виде списка гиперссылок на правления *гунов*. Для каждого правителя пользователь может просматривать текст как по отдельным годам, так и для всего периода. Имеется возможность как отдельного, так и совместного просмотра текстов ЧЦ и ЦЧ. Одно из отличий ZZSTATS от двух предыдущих конкордансов – возможность полного просмотра текста в иероглифах сопровождающегося чтением в транскрипции *пиньинь*, а также реконструкциями доклассических, классических и средневековых чтений иероглифов[4].

В отличие от предыдущих интерактивных конкордансов, в ZZSTATS все иероглифы в текстах сопровождаются гиперссылками к конкордансу, и по этим гиперссылкам пользователь может получить доступ к странице конкорданса для данного иероглифа. Страница конкорданса предоставляет все возможные данные по иероглифу: его возможные чтения, количество вхождений в тексты, а также список фрагментов текстов с данным иероглифом (полным фрагментом является текст сезона или месяца). Сам заглавный иероглиф (node) страницы конкорданса выделяется на ней жирным текстом и квадратными скобками.

Пользователь может выбрать – просмотр только вхождений в ЧЦ, только вхождений в ЦЧ, а также вхождений в оба текста. Кроме того,

для иероглифов с большой частотностью, существует возможность просмотра фрагментов с ограничением длины. Все иероглифы приводимых фрагментов также снабжены гиперссылками. Так же, как и в погодных записях, существует возможность сопровождения фрагментов всеми возможными чтениями иероглифов.

Отличительная черта ZZSTATS – наличие «корзины», или буфера иероглифов. В мексиканском и гонконгском сетевых конкордансах есть возможность поиска по нескольким иероглифам или фразам, но нет возможности накапливать эти иероглифы в буфере для сбора статистики[5]. Страница конкорданса предоставляет возможность просмотра «корзины» и добавления в неё новых иероглифов или удаления иероглифов из неё. Подробнее функциональные особенности «корзины» будут освещены ниже.

Search

Поиск по конкордансу осуществляется в настоящее время двумя способами – как по самому знаку, введённому в соответствующее поле формы поиска, так и по чтению иероглифа в транскрипции *пиньинь*[6]. В первом случае, если иероглиф, по которому осуществляется поиск, присутствует в текстах, то выводится только данный иероглиф, снабжённый гиперссылкой к соответствующей странице конкорданса. Во втором случае, выводится список всех иероглифов с соответствующим чтением. Эти иероглифы также снабжены гиперссылками к своим страницам конкорданса.

Bag

Как говорилось ранее, «корзина» (bag, буфер иероглифов) является основной отличительной чертой ZZSTATS. Интерфейс «корзины» предоставляет возможности для статистического и дистрибуционного анализа групп иероглифов. В настоящее время реализованы четыре функциональных подобласти: просмотр содержания «корзины» (**content**), информация по расстояниям между иероглифами «корзины» (**distances**), анализ пар иероглифов (**bigrams**), и общая статистика (**bag stats**).

Страница содержания корзины (**content**) предоставляет возможность просмотра списка иероглифов в «корзине», а также наличие этих знаков в текстах правления каждого из *гунов*. Это позволяет получить быстрое представление о (совместном) употреблении знаков в соответствующие периоды. Знаки снабжены гиперссылками к соответствующим страницам конкорданса, а имена *гунов* – гиперссылками, позволяющими получить данные о распределении употребления иероглифов по годам правления данного правителя. Это позволяет исследователю быстро определить, встречаются ли иероглифы интересующей его группы в текстах правления различных *гунов*, и, если встречаются, можно ли рассчитывать на их употребление в погодной записи.

Страница информации по расстояниям между иероглифами «корзины» (**distances**), предоставляет исследователю возможность определить характеристики попарного совместного распределения знаков в «корзине» с точки зрения расстояний между иероглифами. (Расстояние измеряется только в пределах сезонного фрагмента.) На данной странице

выводится таблица, в рядах и колонках которой представлено количество фрагментов, где пары любых иероглифов, находящихся в данный момент в «корзине», присутствуют совместно. Эти цифры также снабжены гиперссылками, которые позволяют пользователю получить список таких коллокаций в порядке возрастания расстояния между иероглифами. Цифры расстояний, в свою очередь, также снабжены гиперссылками, которые позволяют увидеть все фрагменты текста, где эти иероглифы встречаются совместно на указанном расстоянии[7].

Страница анализа распределения пар иероглифов (**bigrams**), предоставляет данные для анализа наличия пар иероглифов, соседствующих друг с другом (собственно коллокаций). У неё такая же структура, как и у предыдущей страницы, но в ячейках таблицы находятся данные по количеству коллокаций (если они имеются).

Наконец, страница общей статистики для «корзины» (**bag stats**) аналогична странице содержания «корзины», с тем отличием, что для каждого иероглифа «корзины» предоставляются статистические данные по употребляемости данного знака в различные периоды правления, отдельно по ЧЦ, ЦЧ, и вместе. Эти цифры также снабжены гиперссылками, позволяющими увидеть соответствующие фрагменты.

Stats & Freqs

ZZSTATS предоставляет исследователю разнообразные статистические данные по текстам в целом. В их числе:

1) численное распределение иероглифов по текстам по правлениям *гунов* (**Graph/Gongs**). Строки таблицы снабжены гиперссылками к тексту первых лет правления *гунов*;

2) общие статистические данные по текстам **(Text stats**). Она содержит общее количество иероглифов в текстах, а также количество уникальных иероглифов в текстах. Эти данные снабжены гиперссылками, позволяющими увидеть списки уникальных иероглифов для каждого текста и обоих текстов. Кроме того, здесь можно увидеть количественные данные по иероглифам, которые встречаются только в одном из текстов, снабжёнными гиперссылками к спискам этих иероглифов;

3) список уникальных иероглифов ЧЦ и ЦЧ, в порядке убывания частотности (**All Graphs (freq.)).** Этот список предоставляет исследователю возможность анализа частотности иероглифов (в данной версии текста);

4) список уникальных иероглифов в алфавитном порядке чтений *пиньинь* (**All Graphs (alph.)).** Этот список, в дополнение к странице поиска, облегчает поиск иероглифов по чтениям;

5) частотное распределение отдельных иероглифов по правлениям *гунов* (**Graph Freq**). Строки таблицы снабжены гиперссылками, которые позволяют увидеть распределение иероглифов по частоте для каждого правления года в целом, а также для любого года. Это функция представляет собой мощное средство лингвистического анализа текста;

6) частотное распределение иероглифов по сезонам года и правлениям *гунов* (**Seasons).** Строки таблицы снабжены гиперссылками, позволяющими перейти на соответствующую страницу текста;

7) список биграмм в порядке убывания частнотности (первая сотня) (**Bigrams)**. Этот список важен для выявления наиболее частотных коллокаций пар иероглифов. Он демонстрирует возможности системы в области автоматического анализа текста (синтаксического и POS (части речи) разбора предложений);

8) список триграмм в порядке убывания частотности (первая сотня) (**Trigrams).** Этот список важен для выявления наиболее частотных коллокаций трёх иероглифов. Он демонстрирует возможности системы в области автоматического анализа текста (синтаксического и POS (части речи) разбора предложений).

Заключительная часть проекта – раздел **About Project** – содержит общую информацию об истории и технологии проекта, его истории, а также полезные ссылки и планы развития проекта. В числе этих планов – добавление двух других комментариев, «Гунъян» и «Гулян», а также английских значений иероглифов. Планируется также добавление соответствий словарю рифм. В более отдаленной перспективе – автоматическая аннотация иероглифов текстов частями речи.

Примечания

[1] Первый сетевой конкорданс для ЦЧ был создан в рамках проекта Джона Пейджа и Исабель Гарсия Идальго из университета Мехико [2], при поддержке Брюса Брукса (E. Bruce Brooks), в конце 90-х гг. XX в. (однако оказался доступен пользователям только в 2005 году, примерно тогда же, когда и гонконгская версия). В качестве версии текста для этого конкорданса было избрано издание 1817 г. Жуань Юаня (восходящее к сунскому «*Chongkan Songben Chunqiu Zuozhuan zhushu* 重刊宋本春秋左傳注疏, «Сунское издание комментариев и примечаний к Чуньцю и Цзочжуани»), в версии пятого тома переводов Джеймса Легга, в предпринятом Гонконгским университетом репринтном издании 1962 г. Второй конкорданс, в составе гонконгского проекта CHANT [5], был реализован несколько позже, в середине этого десятилетия. Он также основан на издании Жуань Юаня, с разбиением на главы Yang Bojun's *Chunqiu Zuozhuanzhu* (Beijing, 1990). Существует также несколько сетевых ресурсов с возможностью поиска по тексту (например, известный вебсайт Дональда Старджена [1]), но они не являются конкордансами в полном смысле слова. Насколько нам известно, в настоящее время нет ни одного аннотированного корпуса этого памятника. Проект ZZSTATS доступен в интернете с сентября 2009 г.

[2] В отличие от Мексиканского конкорданса, ZZSTATS не является параллельным текстом; он не содержит ни перевода ЦЧ, ни (в настоящее время) английских переводов иероглифов. Зато он содержит текст ЧЦ, позволяя проводить сравнительный анализ обоих текстов. В отличие от Гонконгского конкорданса, ZZSTATS представляет собой общедоступную систему с более развитыми средствами статистического и текстологического анализа. Ниже различия между этими тремя системами будут описаны более подробно.

[3] Например, система CHANT, несмотря на то, что восходит к той же версии текста, что и мексиканский конкорданс, имеет бóльшую длину текста. Он насчитывает 198 699 знаков для обоих текстов, из них уникальных – 3 320. Мексиканский

проект насчитывает 179 522 иероглифов в «Цзо чжуань». Данные ZZSTATS: ЧЦ – 16 791, ЦЧ – 178 564, оба текста – 195 355, уникальных знаков – 3 252. Меньшая длина ZZSTATS объясняется, в частности, тем, что даты погодных фрагментов не включены в общую длину текста.

[4] Эта возможность существует благодаря любезности Г.С. Старостина, предоставившего нам, в удобной для компьютерной обработки форме, базу данных китайских реконструкций Старлинг [3], созданную С.А. Старостиным, и постоянно развиваемую Г.С. Старостиным. Эта база данных содержит и английские эквиваленты, которые, вероятно, со временем будут добавлены к ZZSTATS.

[5] В ZZSTATS в настоящее время нет возможности поиска по строке или группе иероглифов.

[6] Поиск осуществляется только по значимой форме *пиньинь*. Поиск по частичной форме в настоящее время не производится.

[7] Эта страница пока обладает всеми свойствами обычной страницы конкорданса, т.е., чтения иероглифов в различных формах, и т.д.

Сетевые ресурсы

1. chinese.dsturgeon.net
2. mezcal.colmex.mx/Zuozhuan/Scripts/cuenta.idc
3. starling.rinet.ru
4. zh.wikisource.org/zh-hant/%E6%98%A5%E7%A7%8B%E5%B7%A6%E6%B0%8F%E5%82%B3
5. www.chant.org
6. www.zzstats.com

Sergei Zinin

Toronto

New on-line concordance of the *Chunqiu* and *Zuozhuan*

Abstract

ZZSTATS (Zuo Zhuan Statistics), www.zzstats.com, is an online concordance of the Chunqiu and Zuozhuan. Being the third online concordance of the Zuozhuan, and the second one of the Chunqiu, it is motivated by richer functionality and usability. It provides free public access to the data. Its text view is fully hyperlinked, and for the first time allows retrieving many important statistical features. It also features full text Romanization, with *pinyin*, as well as with Pre-Classic, Classic and Middle Chinese reconstructions (by S. and G. Starostins). One of the most important and distinctive features of the project is the concept of «the Bag», i.e., stored list of characters, that allows the user to retrieve joint distribution statistics for characters in the Bag. The project also features online lists of bigrams and trigrams for texts, which, together with the Bag, provide a better environment for studying collocations in the Chunqiu and Zuozhuan.