

Текст доклада

Слайд 0



Этот доклад анонсирует одноимённую монографию, которая готовится к печати.

Исследование компьютерной структуризации канонических текстов проводилось на примере «Дао дэ цзина» в той его структуризации, которую предложил Артёмий Михайлович Карапетянц.

Мы провели своего рода компьютерную «аудиторскую проверку» этой структуризации, только вместо соответствия законодательным нормам, юридическим и хозяйственным документам у нас соответствие математическим критериям.

Сложность в том, что эти критерии приходилось подбирать и даже «изобретать» применительно к предмету исследования.

О какой структуризации идёт речь?

Слайд 1



Первоначально текст «Дао дэ цзина», как и других канонов, записывался «сплошняком», знаков препинания не было, разбивки тоже.

Потом появилась разбивка на фразы и параграфы.

В итоге имеем четыре уровня текста: иероглиф — фраза — параграф — текст целиком.

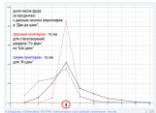
Структуризация параграфов «Дао дэ цзина», которую предложил Спирин и продолжил Карапетянц, вводит дополнительный уровень: параграф разбивается, как правило, на 9 фрагментов, каждый из которых состоит из одной или нескольких фраз.

Эти 9 фрагментов размещаются определённым образом в ячейках 9-клеточной матрицы 3 на 3.

Тем самым, имеется уже пять уровней: иероглиф — фраза — фрагмент — параграф — текст целиком.

Сначала о разбивке на фразы.

Слайд 2



Мы посмотрели три канона: «Дао дэ цзин», «Ши цзин» и «И цзин».

«Дао дэ цзин» взяли в версии Карапетьянца.

Из «Ши цзина» взяли только раздел «Го фын» — «Нравы царств». Фразой считался стих (строка).

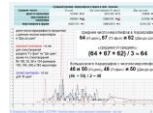
«И цзин» в версии книги Щуцкого под редакцией Кобзева, но без названий гексаграмм и без технических названий черт: в начале шестёрка, в начале девятка, шестёрка вторая, девятка вторая, и т.д.

Средняя длина фразы растёт от «И цзина» к «Дао дэ цзину» от 3,2 до 4,3.

А на графике хорошо видно, что во всех трёх канонах пик приходится на 4: фраз из 4-х иероглифов больше всего.

Теперь разбивка на параграфы.

Слайд 3



Параграфом «И цзина» будем считать текст, относящийся к одной гексаграмме.

Параграфом «Ши цзина» будем считать отдельное стихотворение.

Параграфы «Дао дэ цзина» возьмём в версии Карапетьянца.

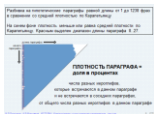
Средние значения размера параграфа в этих трёх канонах очень близки и находятся около числа 64.

На графике по оси абсцисс — размер параграфа, по оси ординат — доля в процентах таких параграфов от общего числа параграфов.

Также близки пики размера параграфа.

В «Дао дэ цзине» это 49 иероглифов, в «Ши цзине» — 48, а в «И цзине» два пика: 46 и 50 иероглифов, среднее значение которых равно 48.

Слайд 4



Одной из побочных целей нашего анализа было установить, насколько «правильна» заданная разбивка текста на параграфы на примере разбивки «Дао дэ цзина» Карапетьянцем.

Эта «правильность» понимается, конечно, не в полном объёме, она основана только на повторении иероглифов.

Мы исходили из гипотезы о том, что при «правильной» разбивке для тройки подряд идущих параграфов параллелизм между фрагментами среднего параграфа должен превышать параллелизм между фрагментами этого параграфа и фрагментами соседних параграфов.

Поскольку параллелизм основан на повторении иероглифов, для каждой такой тройки параграфов мы посчитали долю в процентах числа разных иероглифов в среднем параграфе, которых нет в соседних параграфах, от общего числа разных иероглифов в среднем параграфе.

Такую оценку мы называли *плотностью* среднего параграфа, посчитали её для каждого параграфа и взяли среднее значение.

Но с чем сравнить эту среднюю плотность?

Большая она или маленькая?

Сначала мы провели очень грубое сравнение: разбивали текст на гипотетические параграфы равной длины от одной до 1238-и фраз (столько во всём тексте).

Результат показан на рисунке.

Синим цветом отмечены случаи, когда средняя плотность по Карапетьянцу больше или равна плотности гипотетического параграфа.

Ну, мы видим, что в основном всё окрашено в синий цвет.

Красными линиями выделен диапазон длин гипотетических параграфов от 8 до 27 — столько фраз в параграфах по Карапетьянцу.

Здесь тоже преобладает синий цвет, хотя есть и белый.

► Для более глубокого анализа мы стали рассматривать тройки гипотетических параграфов

РАЗНОЙ длины в диапазоне от 8 до 27 фраз, но с тем же началом среднего параграфа, как у Карапетьянца.

Оказалось, что в среднем плотность параграфа Карапетьянца составляет 83% от максимально возможной.

Это тоже хороший результат, если принять во внимание, что мы не учитывали синтаксис и семантику.

Слайд 5



Итак, текст разбит на параграфы, параграф разбит, как правило, на 9 фрагментов, а фрагмент состоит из целого числа фраз.

Как же размещаются эти 9 фрагментов в 9-клеточной матрице 3 на 3?

Нужно выбрать размещение с максимальным параллелизмом.

Параллелизм считается между фрагментами в одном ряду: в одной строке или в одном столбце.

Вообще, по Спирину параллелизм основывается, главным образом, на таких формальных параметрах, как длины фрагментов, совпадение иероглифов и, возможно, их позиций и рифма.

Далее учитывается синтаксис и, в последнюю очередь, семантика.

Мы ограничились только формальными параметрами, синтаксис и семантику не учитывали.

На слайде размещение 1-го параграфа Даодэцзина.

Слева чёрными стрелками показан порядок чтения текста.

У одинаковых иероглифов одинаковый цвет фона, а у рифмующихся иероглифов одинаковый цвет шрифта (красный, синий или зелёный).

Справа цветными квадратами изображен цветной фон, а цветными кружками — рифма.

Стрелками показаны все параллелизмы такого рода.

Слайд 6



Критерий	Описание
Равенство длин	Длины фрагментов равны
Разность длин	Длины фрагментов различаются
Повторение иероглифов	Иероглифы повторяются в обоих фрагментах
Подпоследовательности	Иероглифы образуют подпоследовательности в обоих фрагментах
Расстояния между иероглифами	Расстояния между иероглифами совпадают
Транспозиция	Иероглифы соседних позиций обмениваются местами
Удаление/вставка/замена	Иероглифы удаляются, вставляются или заменяются

На основе формальных параметров можно выделить много критериев подобия (параллелизма) двух фрагментов текста.

Мы использовали 19 таких критериев.

Они на слайде. Есть 4 группы критериев.

Первую группу образуют два критерия основанные на длине фрагмента: равенство длин и разность длин.

Потом большая группа критериев, основанных на повторении иероглифов, а именно: на общих подпоследовательностях иероглифов в двух сравниваемых фрагментах.

В свою очередь эта группа делится на 4 подгруппы в зависимости от того, учитываются или не учитываются расстояния между иероглифами подпоследовательности, до её первого иероглифа и после её последнего иероглифа.

То есть: иероглифы соседние в подпоследовательности в самом фрагменте могут

быть не соседними, между ними может быть то или иное расстояние.

Также первый иероглиф подпоследовательности не обязательно первый во фрагменте, а последний иероглиф подпоследовательности не обязательно последний во фрагменте.

► Объясню это на примере двух строк букв.

На слайде показаны наибольшие общие подпоследовательности слов МЕТРИКА и ЧЕТВЕРТАК без учёта расстояний и с учётом тех или иных расстояний.

Третья группа критериев определяется через минимальное число элементарных операций, которые нужно сделать, чтобы один фрагмент превратить в другой фрагмент. Операции такие: удаление одного иероглифа, вставка одного иероглифа и замена одного иероглифа другим иероглифом.

Дополнительно может быть транспозиция соседних иероглифов.

Четвёртая группа состоит из одного критерия — это рифма.

Важно отметить, что ни один из этих критериев не может претендовать на лидерство.

У Артёма Михайловича уже не спросишь, какие критерии ему нравятся.

А Спирин писал, что без формального подхода к тексту получается не столько наука, сколько разновидность эссеистики, в которой главную роль играют интуиция и эрудиция, а иногда и просто личный авторитет и другие случайные обстоятельства.

Ирония в том, что эти слова можно отнести и к той структуризации, которую предложили Спирин и затем Карапетьянц.

Поскольку нет формальных причин предпочесть один критерий другим или выстроить их в некую иерархию, мы провели многокритериальную оптимизацию, которая даёт, вообще говоря, не единственное размещение, а несколько «самых параллельных» размещений.

Суть её в следующем.

Для каждого размещения и каждого критерия вычисляется число параллелизма для каждой пары фрагментов одном ряду (строке или столбце).

Эти числа суммируются для данного размещения и данного критерия.

Для данного размещения получается вектор из 19 таких сумм — по числу критериев.

Размещение считается максимально параллельным, если любое другое размещение хотя бы по одному критерию меньше или по всем критериям такое же.

Ну, математики говорят, что на векторах не линейный, а частичный порядок.

Слайд 7



На слайде слева 14 канонических размещений, выделенных Спириным.

В нашей работе мы сформулировали свойства таких размещений и доказали их необходимость и достаточность.

Некоторые из этих свойств были указаны ещё Спириным, другие, видимо, остались в его интуиции.

На самом деле с точностью до эквивалентности канонических размещений только 7.

Остальные получаются симметрией матрицы относительно главной диагонали.

Иными словами, отличие в том, двигаемся мы от 1-го ко 2-му фрагменту по горизонтали или по вертикали.

Справа — таблица, которая вычислена для 1-го параграфа «Дао дэ цзина» в двух вариантах: без учёта деления фрагмента на фразы и с учётом такого деления.

Там немного разные алгоритмы.

Строка таблицы соответствует критерию, здесь указаны только их номера.

Столбец соответствует каноническому размещению, это как раз вектор чисел, о котором я говорил.

Вот эти столбцы и сравниваются между собой при многокритериальной оптимизации.

Числа максимальные в строке выделены красным фоном, остальные ячейки фиолетовые.

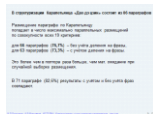
Исключение сделано только для столбца, соответствующего размещению Карапетьянца. Здесь вместо красного цвета оранжевый, а вместо фиолетового жёлтый. Есть ещё зелёный цвет — это когда число максимально в строке и в этой строке нет других таких же чисел.

Максимально параллельные размещения, т.е. столбцы, отмечены красным или оранжевым фоном ячейки заголовка.

Оранжевый — это для размещения Карапетьянца, если оно максимальное.

Для 1-го параграфа максимальны два размещения и одно из них — размещение Карапетьянца.

Слайд 8



Такое попадание размещения Карапетьянца в число максимальных имеет место для 66 или 63 параграфов в вариантах без учёта и с учётом деления на фразы.

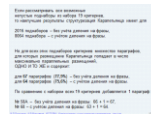
Это составляет, соответственно, 77 и 73 процента от общего числа параграфов.

Мы считаем, что это хороший результат, если учесть, что мы игнорировали синтаксис и семантику.

Это более чем в полтора раза больше, чем математическое ожидание при случайной выборке размещения.

В подавляющем большинстве параграфов результаты с учётом и без учёта деления на фразы совпадают.

Слайд 9



Потом мы подумали: а что если какие-то критерии противоречат интуиции Карапетьянца, и если их убрать, то результат будет ещё лучше.

В общем, так и оказалось.

Мы провели вычисления не только для набора всех 19 критериев, но и для каждого непустого поднабора критериев.

Таких поднаборов больше полмиллиона.

Так были выявлены самые лучшие для Карапетьянца поднаборы критериев.

Набор из всех 19-и критериев оказался не самым лучшим.

Таких лучших поднаборов критериев оказалось много — несколько тысяч.

Но интересно и неожиданно то, что для всех этих поднаборов сами параграфы, для которых размещение Карапетьянца среди лучших, одни и те же.

Только один параграф добавляется по сравнению с набором всех 19-и параметров.

Но почему так много наилучших поднаборов критериев? Как они устроены?

Слайд 10



А устроены они очень регулярно.

Вот на слайде это показано.

Без учёта деления на фразы два критерия — на белом фоне со словом НЕТ —оказались вредными для Карапетьянца.

5 критериев — на жёлтом фоне — наоборот, обязательно нужно использовать.

5 критериев — на зелёном фоне — не влияют: их можно использовать, а можно и не использовать.

Для двух пар критериев — на синем фоне — нужно использовать хотя бы один из двух критериев каждой пары.

Наконец, на фиолетовом фоне — три критерия, из которых нужно использовать хотя бы один.

Слайд 11



На этом слайде то же самое, но для случая, когда учитывается деление текста на фразы.

В заключение хочу отметить, что разработанные методы, алгоритмы и компьютерные программы могут применяться не только к структуризации Карапетьянца, и не только к тексту «Дао дэ цзина».

Кроме того, кое-что можно использовать для решения задач, весьма далёких от текстологических исследований.

Это у меня последний слайд — одиннадцатый.

Слайд 12



Спасибо за внимание!