

МОДЕЛИРОВАНИЕ СЛОВАРЕЙ ОПИСАНИЙ ДОКЕМБРИЙСКИХ АКРИТАРХОВ

А.А. Огай

Санкт-Петербургский государственный университет

Санкт-Петербург

Важной проблемой в компьютерной лингвистике является проблема отображения полученных ею данных. Стюарт К. Кард определял визуализацию информации в целом (ВИ) как «использование интерактивных и поддерживаемых компьютерами репрезентаций абстрактных данных для улучшения когниции». Методы ВИ включают в себя различные графики, диаграммы и другие, менее ортодоксальные средства [1-4].

Связана с лингвистикой одна из областей ВИ: визуализация языковой информации, или т.н. LInfoVis, либо ВЯИ — применение принципов информационной визуализации для отображения любого вида информации, связанного с языком и его использованием. ВЯИ отличается от других ветвей ВИ из-за особенностей языковой информации и сложностей, связанных с ее измерением.

Тем не менее, ВЯИ может принести значительную пользу для научной работы, так как позволяет организовывать, отмечать и сравнивать эту информацию и обнаруживать в ней тренды и исключения из этих трендов — «создавать и открывать идеи». Помимо этого, ВЯИ за счет своей наглядности может служить и для визуально яркого отображения уже известных данных.

Однако визуализация также сопряжена и с рядом особых, характерных именно ей проблем. В частности, при создании средств ВЯИ важно, чтобы визуальная привлекательность модели («красивые картинки», как выразился Роберт Косара) не превалировала над её информативностью. В частности, «облака слов» (word clouds) — это однозначно «красивые картинки», однако не содержат большого количества информации, кроме частотности.

Из-за этого развитие ВЯИ невозможно без соблюдения определенных принципов, призванных обеспечить эффективность используемых методов визуализации. Ученые Европейской академии Больцано выделяют такие принципы: визуальная ясность, прозрачность содержимого, организация данных и интерактивность [5].

Помимо этих принципов, автор также добавляет еще один: автоматизация. Желательно, чтобы используемый материал не нужно было подвергать значительной обработке перед анализом. Человек должен работать как можно меньше: все исправления исходного материала должны происходить непосредственно внутри программы.

В данном докладе методы визуализации будут рассмотрены на примере авторских частотных словарей. Авторская и частотная лексикографии исторически были близко связанными: причиной тому является тот факт, что идиолект, то есть совокупность уникальных особенностей речи отдельного носителя языка, проявляется зачастую именно в частотности определенных слов или конкордансов. По этой причине есть смысл также коротко обсудить эти ветви лексикографии подробнее.

Частотные словари впервые начинают создаваться в Европе в эпоху позднего Средневековья: создающиеся в это время конкордансы подсчитывали в том числе и частоты появления каждого слова в тексте (авторском либо библейском). К XIX веку частотные словари распространяются за пределы стран западной Европы и становятся глобальным явлением, а с изобретением компьютера в середине XX века доступ к мощным вычислительным машинам и крупным корпусам дает этой дисциплине новый рывок за счет упрощения процесса сбора языковой информации.

Авторские частотные словари, рассматриваемые в рамках доклада, документируют научные работы по исследованию докембрийских акритархов — древних микроорганизмов, существовавших примерно 1 миллиард лет назад. Так как эти организмы существовали еще до так называемого «кембрийского взрыва», когда появилась скелетная жизнь, их останки находятся гораздо реже останков животных и растений более поздних эпох. Как следствие, существование этих микроорганизмов долгое время ставилось под сомнение и остается предметом многих вопросов до сих пор.

Акритархи изучались в СССР несколькими учеными, которые, однако, не были профессионалами в области биологии. Среди них Борис Васильевич Тимофеев, изначально изучавший торф на территории Ленинградской области, и геолог Нина Андреевна Волкова. Работы этих двух ученых в разные периоды будут исследоваться в рамках этой работы. Из-за отсутствия у этих двух исследователей профессионального палеонтологического образования их работам присущ ряд непрофессиональных черт (например, использование косвенных падежей в определениях видов, в таксономии неприемлемое).

Визуализация частотных словарей работ этих двух ученых дает возможность сравнить их исследования и увидеть различия в данных, ими полученных, а также позволит оценить работу каждого из них в

отдельности (в частности, как уже было сказано, по профессиональности). В пределах данного доклада будут показана визуализация в формате трехмерных диаграмм.

Таблица. Частотные словари

1	ОБОЛОЧКИ	87	86	3,798
2	ДО	82	168	3,759
3	ОТ	77	245	3,361
4	И	67	312	2,924
5	ИЛИ	67	379	2,924

Работы двух ученых разделены по тематике: надвидовые и видовые таксоны. Кроме того, работы Тимофеева разделены также по периоду (первый и второй вместе с третьим): в сумме мы имеем 6 различающихся друг от друга частотных словарей, которые возможно объединять по общим характеристикам (например, все надвидовые словари).

Сама программа состоит из "анализатора" и "визуализатора". Исходный материал в виде текстового файла с данными частотных словарей проходит обработку, в ходе которой также складываются некоторые отдельные виды, после чего эти обработанные данные преобразовываются в диаграммы.

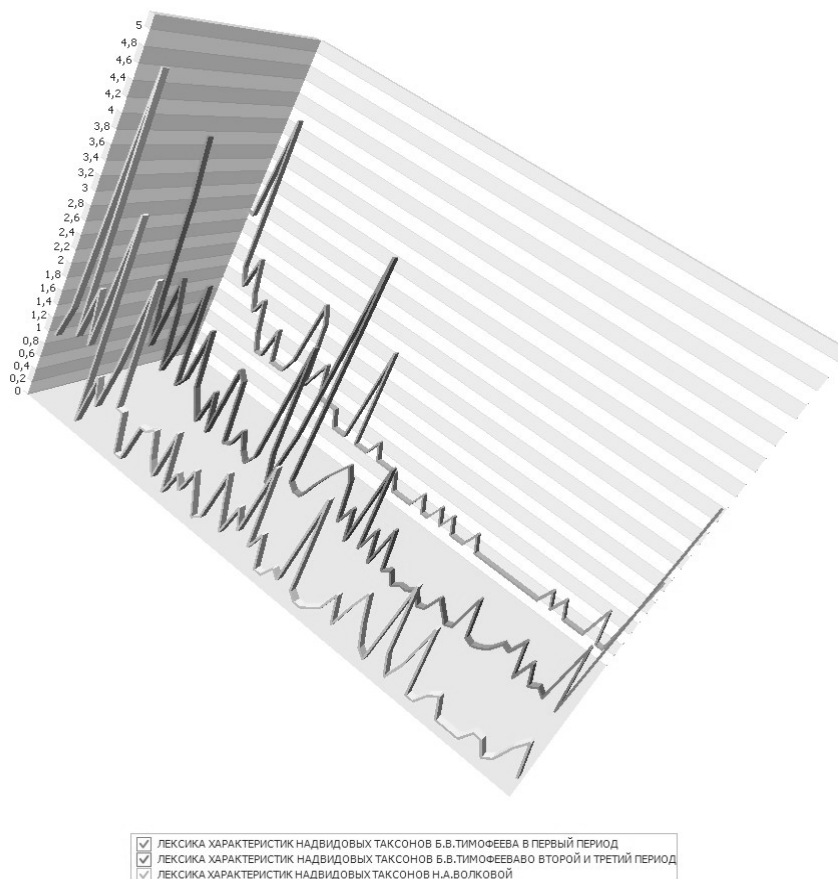


Рисунок. Один из ранних вариантов диаграммы

Диаграммы являются видоизменяемыми. Помимо простой диаграммы, показывающей частоты, взятые из словарей, возможно также сравнивать самые сильно различающиеся и самые однородные их части. Кроме того, можно будет использовать либо не использовать лемматизацию. Вариант, не использующий лемматизацию, полезен для выделения косвенных падежей, которые, как было упомянуто выше, являются важными для статистики.

Уже имеющиеся диаграммы показали ряд различий частотных словарей друг от друга и тенденций. В частности, словарный запас расширяется от надвидовых словарей к видовым и с течением времени, среди самых распространенных слов заметно наибольшее колебание в частоте; в видовых словарях гораздо чаще

встречаются числа и меры измерения. Численные значения в работах Н. А. Волковой встречаются значительно реже, в то время, как в работах Б. В. Тимофеева их частотность, напротив, увеличивается во втором и третьем периодах.

Наработки, предоставленные в докладе, не будут ограничиваться исключительно лишь работами, посвященными акритархам, и могут быть использованы (в дальнейшем) и для другой научной литературы.

ЛИТЕРАТУРА

1. Алексеев П. М. Частотные словари. Учебное пособие, Санкт-Петербург, 2001.
2. Ковригина Л. Ю. Негауссовое моделирование лексико-статистической структуры вариативного текста (на примере «Сказания о Мамаевом побоище»), Санкт-Петербург, 2014.
3. Фуфаев В. В. Структурно-топологическая устойчивость динамики ценозов // Кибернетические системы ценозов: Синтез и управление. МОИП. IX чтения памяти А.А. Ляпунова. М., 1991.
4. Chris Culy, Verena Lyding: Visualizing Linguistic Data: From Principles to Toolkits for Doing it Yourself. AVML Conference, 2012.
5. Chris Culy, Verena Lyding. Visualisation of linguistic information, Bolzano, 2010.