МНОГОМЕРНЫЙ АНАЛИЗ ЯЗЫКОВЫХ И РЕЧЕВЫХ ДАННЫХ

Темиров Б.

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Петрова Н.Е. – к.филол.н., доцент

Статья посвящена исследованию структурной организации системы языка, выявление связей и зависимостей между его элементами и признаками. Рассматривается методика многомерного анализа.

В число важнейших задач лингвистики входит исследование структурной организации системы языка, выявление связей и зависимостей между его элементами и признаками. Система с точки зрения ее организации характеризуется как некоторым набором элементов определенной субстанции (субстанциальный аспект), так и схемой отношений своих элементов (реляционный аспект).

Одним из основных направлений анализа структуры объекта является построение классификации его элементов (единиц и признаков). Осуществляя классификацию, мы создаем конструкт, моделирующий структуру исходных данных. Такой подход рассматривается как самый простой и естественный способ анализа информации, отражающий логику эмпирического исследования.

Понятие классификации и связанные с ней понятия таксономии, систематики и др. являлись предметом неоднократного рассмотрения в работах как общетеоретического плана, так и исследований, направленных на решение задач конкретных областей знаний (работы С. В. Мейна, Ю. А. Шрейдера, С. С. Розова, А. А. Любищива; Р. Р. Сокала П. Х. А. Снита и др.).

Классификационное упорядочение данных занимает важное место в любом лингвистическом исследовании, направленном на анализ конкретного языкового материала. В результате получены детальные описания и общие схемы, отражающие как отдельные участки языка, так и целые уровни и планы языковой системы. Вместе с тем следует признать, что в большинстве лингвистических классификаций проявляются следующие тенденции, реализующиеся в них в различной степени:

1. Во многих классификациях не выдерживается принцип системности и последовательности анализа. Это проявляется в том, что классификационная схема признаков не является единой для сопоставляемых классов и не соотносится со всеми объектами, включенными в исследования, т.е. не применяются ко всему исследуемому материалу. Иными словами, имеет место такое положение, когда к одной группе единиц применяются признаки одного вида, а к другой группе – признаки, выделенные на иных принципах. Вполне естественно, что в таких исследованиях полных списков признаков не приводится, а сам анализ приобретает вид комментария. Кроме того, следует отметить, что часто не учитывается общая представленность признака на всей совокупности материала, из-за чего трудно судить о диагностической силе этого признака (не ясно, имеется ли он в других классах и насколько часто он там встречается).

2. Большинство лингвистических классификаций не учитывает реляционного аспекта систематики – соотношений и связей, существующих между признаками и элементами языковой системы.

По мнению Б. Г. Миркина, проанализировавшего большое количество классификационных задач в биологии, социологии и ряде др. таксономических наук, все они могут быть сведены к двум основным типам, соответствующим двум основным средствам обогащения теоретических представлений о системе (конструирование новых категорий, отражающих существенные стороны системы, и установление новых закономерностей, связывающих различные категории): конструирование признаков и описание одних признаков значениями других [1].

Значение реляционного аспекта в систематике отражается уже в самом определении «естественности» классификационной схемы (ее научной значимости, адекватности описываемых явлений). Считается, что классификация является естественной в том случае, если она позволяет выводить или предсказывать наличие у элемента признаков, исходя из его других базовых характеристик, определяющих место этого элемента в системе [2].

Мерой естественности классификационной системы в этом случае признается ее прогностическая (импликационная) сила, количество существующих связей между элементами, которые она отражает.

3. В абсолютном большинстве случаев систематизация материала основывается на фактах, выделяемых на каком-либо одном лингвистическом уровне.

Признание уровней структуры в качестве имманентной сущности языка сделало правомерным и оправданным изучение изолированных подсистем без учета их внешних, межуровневых связей. Однако по мере накопления информации об отдельных уровнях естественно поставить вопрос о переходе от одноуровневого анализа к многоуровневой классификации единиц. Такой подход помогает не только выяснить характер взаимодействия единиц различных подсистем, но и взаимосвязей между подсистемами в целом. С

другой стороны, учет межуровневых соотношений позволяет в целом ряде случаев по-новому построить схему отношений и в рамках одной, отдельно взятой подсистемы.

4. Многим лингвистическим классификациям в большой степени свойственен описательный метод упорядочения материала. Он заключается в отсутствии математической обработки данных и, как следствие этого, проверке степени достоверности результатов, справедливости получаемых выводов для широкого лингвистического материала.

5. В тех случаях, когда исследователь обосновывает свои выводы не отдельными примерами, а обращается к систематическому анализу материала, в его распоряжении, как правило, оказывается большое количество данных. Обилие статистической информации делает невозможным ее 45 использование непосредственно как базы классификационного анализа единиц – группировки признаков и объектов. Закономерности структурной организации, отражаемые в эмпирических таблицах, распыляются на множество фиксируемых связей. Огромное число индивидуальных наблюдений не позволяет выделить наиболее существенные тенденции. Причем любая попытка углубить исследование вызывает еще большее возрастание признакового пространства, что может привести к тому, что общая картина станет вообще необозримой. Поэтому возникает необходимость обратного сведения этой многообразной информации к небольшому количеству обобщающих выводов [3; 4; 5].

Указанная цель по «сжатию» представленной информации для большого количества разноуровневых признаков, определяющих пространство исследования, может быть достигнуто путем проведения многомерного анализа данных.

В область многомерных задач исследование вступает в том случае, если начинает учитываться совместная вариация как минимум двух признаков для нескольких объектов. Многомерные данные, таким образом, задают значения нескольких переменных для нескольких объектов, то есть Xij, где i – соответствует объекту, а j – переменной.

В рамках коллективной темы, проводимой в Смоленском государственном университете по многомерному анализу языковых и речевых данных, используется подход, с привлечением большого набора (от десятков до сотен) разноуровневых признаков, включающих фонетические, морфемные и деривационные, морфологические, синтаксические, семантические характеристики, а также (при исследовании стиля авторов на стихотворном речевом материале) ритмометрические, строфические, рифменные и др. Они выявляются у языковых и речевых элементов на больших массивах текстов или репрезентативных языковых выборках с последующим многомерным анализом полученных баз данных. В качестве основных математических методов анализа используются многомерный дискриминантный анализ, многомерный регрессионный анализ, различные виды кластерного анализа, факторный анализ [6; 7; 8].

Следует отметить, что применяемый нами подход в целом соответствует методике анализа стилеметрических исследований.

Вместе с тем, в дальнейшем предстоит, по-видимому, ещё более углубленно изучать как специфику применения различных конкретных процедур многомерного анализа на языковом материале (выбор исходного списка признаков в качественном и количественном аспектах, алгоритмы сокращения признакового пространства, соотношение различных видов группировок), так и способы интерпретации получаемых разбиений, а также саму стратегию многомерного исследования в целом.

Список используемых источников

1. Миркин, Б.Г. Анализ качественных признаков и структур / Б.Г. Миркин. – М.: Статистика, 1980. 8 Тулдава, Ю. Опыт квантитативного анализа художественного стиля / Ю. Тулдава // Studia metrica et poetica. Tartu, 1976. – С. 122–141.

2. Дюран, Б. Кластерный анализ / Б. Дюран, П. Одел. – М. : Статистика, 1977. Баевский, В. С. Лингвистические, математические, семиотические и компьютерные модели в истории и теории литературы / В.С. Баевский. – М.: Языки славянской культуры, 2001.

3. Тулдава, Ю. Опыт квантитативного анализа художественного стиля / Ю. Тулдава // Studia metrica et poetica. Tartu, 1976. – С. 122–141.

4. Кучер, И.Н. Структурно-семантические признаки именных образных моделей в английском языке (на материале образной системы А. Теннисона). Автореф. дис. … канд. филол. наук / И. Н. Кучер; Смоленский государственный университет. – Смоленск, 2006.

5. Марусенко, М.А. Атрибуция анонимных и псевдонимных литературных произведений методами теории распознавания образов / М. А. Марусенко. – Л. : Издательство Ленинградского ун-та, 1990.

6. Мартыненко, Г. Я. Основы стилеметрии / Г. Я. Мартыненко. – Л. : Изд-во Ленинградского ун-та, 1988.

7. Андреев, В. С. Динамика стиля Э. По (На материале лирики) / В. С. Андреев // Известия Российского государственного педагогического университета им. А. И. Герцена. – СПб, 2008. – № 11 (72). – С. 168–174.

8. Забродин, В. Ю. О критериях естественности классификаций / В. Ю. Забродин // Научно-техническая информация. ВИНИТИ. Сер.2. № 8. – 1981. – С. 22–24.