Кластерный анализ и его применение в лингвистике.

Дефиниция термина «кластер»

Для определения понятия «кластерный анализ» сначала необходимо разобраться с дефиницией термина «кластер». В библиотечном деле существует следующее определение «кластера»: «коллекция слов и выражений, которые ассоциативно связаны и не обязательно должны иметь общие семантические отношения, которые могут быть формализованы» [Informationsordbogen] С точки зрения статистики под «кластерами» следует понимать «однородные подгруппы, формируемые методом, который минимизирует дисперсию внутри групп (кластеров) и максимизирует дисперсию между группами» [Браун] Существует и экономическое определение «кластера», понимающее термин как некоторую политико-экономическую агломерацию, но в контексте данного исследования это малорелевантно.

Кластером в лингвистике называют «совокупность лексем, различной частеречной принадлежности», которые репрезентируют фрагмент картины мира. [Хроленко 1995, с.73] Также «кластер» - это «сегмент информационного поля» и «лексикографическое описание всех входящих в кластер лексем с параллельным установлением всех связей каждого слова с остальными словами, представляющими один и тот же фрагмент фольклорной картины мира» [там же]. Н.Г. Комлев в «Словаре иностранных слов» так толкует термин «кластер»: «лингв. последовательность, цепочка языковых элементов, которыми могут быть звуки <…> или части речи <…>» [Комлев, с.57]

Регулярно используется термин в лингвокультурологии и фольклористике: см. «Кластер « пища» в русской эпической традиции с точки зрения территориальной специфики» [Праведников]. «Кластер» используют как «часть языка»: «Статья посвящена рассмотрению проблем перевода звукоизобразительной лексики турецкого языка средствами английского языка и освещает основные способы перевода данного кластера лексики» [Краснова, с.1]

 Исследователи отмечают, что в современной литературе сосуществует синонимичное использование терминов «кластер» и «класс» и их дифференцирование. Датский ученый Birger Hjørland, который специализируется в библиотечном деле и организации информации, отмечает, что точка зрения о необходимости разделения терминов «класс» и «кластер» восходит еще к Людвигу Витгенштейну [Hjørland], который и предложил термин «кластер» в «Философских исследованиях». Его идея «кластера» или «семейного сходства» заключалась в следующем: «признании особого вида обобщения группы предметов, сходных в одном, но не сходных в другом отношении, что не позволяет обобщать их через род и видовое отличие согласно канонам логики». [Микишина] Витгенштейн приводит в пример идею игры: чем больше человек думает, тем яснее становится тот факт, что составить список характеристик, которые необходимы и достаточны для того, чтобы описать «игру», очень сложно. Так, после нескольких неуспешных попыток, он оставил идею описать «язык» и ввел термин «семейное сходство».

Еще один термин, функционирующий параллельно с термином «кластер» - термин «поле». Александр Хроленко так объясняет разницу между этими двумя понятиями: поле является проявлением системного характера языка, тогда как кластер является сегментом информационного пространства. [Хроленко 1995, 74]

Кластерный анализ и его задачи

Термин «кластерный анализ» впервые был использован американским психологом Робертом Трионом в одноименной работе еще в 1930 году [Корлисс, с.1]. Несмотря на это, термины «кластер» и «кластерный анализ» воспринимаются носителями языка как новые, что отмечает Александр Хроленко, который провел корпусный анализ употребления лексемы «кластер»: «большинство авторов, использующих этот термин, обращает внимание на его новизну» [Хроленко, 2016, с. 106 ]

Кластерный анализ включает в себя множество различных алгоритмов классификаций, цель которых, организовать информацию в кластеры. Важно помнить, что кластерный анализ не является сам по себе определенным алгоритмом, а есть задача, которую нужно решить. В своей работе «Скудность линейной иерархии» Марк Эрешефски отмечает, что кластерный анализ есть один из трех видов классификации объектов окружающего мира, наряду с эссенциализмом и исторической классификацией.

В лингвистике кластерный принцип описания подразумевает помимо анализа входящих в этот кластер единиц, еще и анализ отношений внутри них. Это могут быть связи разных уровней: от логического (парадигматического и синтагматического, например) до словообразовательных и фонетических связей.

Ф. Браун выделяет следующие шаги кластерного анализа[Браун]:

1. Выбор меры и произведение необходимых измерений, критериев или сущностей, подлежащих классификации
2. Задание меры сходства
3. Формулирование правил для определения порядка формирования кластеров
4. Применение правил для формирования кластеров

Следует отметить,что третий пункт вызывает вопросы, так как отличительной чертой кластеризации как метода классификации является отсутствие заданных классов. Кластеризация документов – это задача информационного поиска. В отличие от категоризации текстов, он не включает в себя заранее определённых категорий или обучающей выборки. Кластеры и отношения между ними «автоматически извлекаются из документов, и документы последовательно прикрепляются к этим кластерам» [Голуб, с. 52-53] Марк Эрешефски представляет кластерный анализ как метод классификации. Он считает, что «все формы кластерного анализа строятся на двух допущениях: члены таксономической группы должны иметь кластер общих черт, и эти черты не могут появляться во всех или только в одном члене этой группы». [Ereshefsky, с. 15]

В своей работе «Кластерный подход в лингвистическом анализе»[Нургалиева] Н.Х. Нургалиева выделяет четыре основные задачи кластерного анализа:

1. Разработка типологии или классификации
2. Исследование полезных концептуальных схем группирования объектов
3. Представление гипотез на основе исследованных данных
4. Проверка гипотез или исследований для определения того, действительно ли типы (группы), выделенные тем или иным способом, присутствуют в имеющихся данных

Все методы кластерного анализа можно разделить на «жесткий», четкий кластерный анализ, когда каждый объект либо принадлежит кластеру, либо нет, и на «мягкий», нечеткий кластерный анализ, когда каждый объект принадлежит какой-то группе с определенной степенью вероятности.

Методы кластерного анализа также делятся на иерархические и неиерархические. Иерархические методы подразумевают наличие вложенных групп, в отличие от методов неиерархических. Хургалиева отмечает, что иерархический метод«представляется наиболее подходящим для решения лингвистических задач»[Хургалиева, с.1], так как он позволяет увидеть и проанализировать структуру изучаемого явления.

Ход кластерного анализа

Алгоритмы кластерного анализа требуют на вход матрицу n\*m, где n – рассматриваемые объекты, а m – признаки. Пошагово кластерный анализ можно описать следующим образом:

1. Все существующие объекты разделяются на отедльные кластеры, такм образом в каждом кластере есть только один элемент
2. Заданный алгоритм объединяет в кластер элементы, с наименьшим расстоянием

Последний шаг итерируется до тех пор, пока не останется один кластер, который содержит все элементы. Результатом такого анализа становится дендрограмма.

Дендрограмма – ветвящаяся диаграмма, которая отражает отношения сходства между группами сущностей.