Кластерный анализ и его применение в лингвистике

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, это 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. 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.

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]

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».

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]

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]

Кластером в лингвистике называют «совокупность лексем, различной частеречной принадлежности», которые репрезентируют фрагмент картины мира. [Хроленко 1995, 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, 57]

В лингвистике кластерный принцип описания подразумевает помимо анализа всходящих 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.

.**Кластерный подход** – это лексикографическое описание всех входящих в кластер лексем с параллельным установлением всех связей каждого слова с остальными словами, представляющими один и тот же фрагмен картины мира.У кластерного подхода обнаруживаются значительные эвристические возможности. Он помогает толковать слова, находящиеся за пределами литературного языка и не отмеченные диалектологическими словарями, интерпретировать т.н. «темные» слова, создает благодатное исследовательское поле для специалистов по морфологии и словообразованию и т.д.