
Короткие сообщения

СВЯЗЬ СТЕПЕНИ КЛАСТЕРИЗАЦИИ ЛЕКСИКО-СЕМАНТИЧЕСКОЙ КАТЕГОРИИ СО СКОРОСТЬЮ ПЕРЕЧИСЛЕНИЯ ВХОДЯЩИХ В НЕЕ ОБЪЕКТОВ

А.А. ГРИГОРЬЕВ^a

^a ФГБУН «Институт психологии РАН», 129366, Москва, ул. Ярославская, д. 13, к. 1

Резюме

В ситуации, когда испытуемых просят перечислить названия объектов, принадлежащих к лексико-семантическим категориям (члены категорий), члены одних категорий перечисляются быстрее, чем члены других. Результаты исследований процесса выполнения данной задачи позволяют предположить, что фактором данного различия лексико-семантических категорий может быть степень кластеризации их членов: категории, члены которых лучше сгруппированы, имеют в этом отношении преимущество. Для проверки этого предположения использовались категориальные нормы для русского языка, содержащие данные о среднем количестве членов категорий, перечисляемых испытуемыми за единицу времени, и о кластерах, содержащихся в категориях и влияющих на ответы испытуемых. Анализировались данные по 22 категориям («драгоценный камень», «четвероногое животное» и т.д.), для каждой из них было подсчитано число кластеров с парной близостью больше трех, данный показатель выступал как степень кластеризации. Как оказалось, по этому показателю категории значительно варьируют: в то время как у двух категорий («драгоценный камень», «инструмент плотника») он был равен нулю, у четырех других («четвероногое животное», «предмет одежды», «родственник», «птица») он был больше 10. Корреляция степени кластеризации со средним числом перечисляемых членов категории составила 0.838; $p < 0.001$. Этот результат согласуется с предположением о роли степени кластеризации членов лексико-семантической категории как фактора скорости перечисления входящих в нее объектов. Однако такими факторами могут явиться и другие различия структур лексико-семантических категорий, в частности, может играть роль тип этой структуры.

Ключевые слова: семантическая память, презентация, лексико-семантическая категория, категориальные нормы, кластеризация.

Введение

Результаты ряда исследований хранения и извлечения информации из семантической памяти в норме и патологии (см., напр.: Warrington, Shallice, 1984; Pilgrim et al., 2005) выявили различия между стимулами, принадлежащими к разным лексико-семантическим категориям. В частности, было показано, что испытуемые быстрее осуществляют категоризацию названий живых объектов, чем неживых (Pilgrim et al., 2005).

В настоящей работе рассматривается диссоциация таких лексико-семантических категорий, как «фрукты», «инструменты плотника» и т.д., в ситуации, когда испытуемых просят перечислять названия объектов, принадлежащих к одной из таких категорий, в течение некоторого времени (*category fluency task*). В отечественной литературе использовались термины «семантическая вербальная беглость» (Алфимова, 2010) и «семантическая беглость» (Дроздова и др., 2015). Эта диссоциация состоит в том, что названия объектов, принадлежащих к одним категориям, перечисляются быстрее, чем названия объектов, принадлежащих к другим категориям, что отражает, надо полагать, различия в скорости извлечения из семантической памяти. На данное различие лексико-семантических категорий обращалось мало внимания, исследователей больше интересовало сопоставление результатов, полученных на отдельных категориях (как правило, «животные»), с данными, полученными при использовании другой задачи, когда испытуемых просят перечислить слова, начинающиеся на определенную букву (см.: Bokat, Goldberg, 2003).

Различие лексико-семантических категорий в скорости извлечения релевантной информации выступает очевидным образом в категориальных нормах. Категориальные нормы — это списки слов, которые носители данного языка называют (точнее, записывают или вводят в компьютер), когда их просят перечислить названия объектов, входящих в лексико-семантические категории (члены категорий). В категориальных нормах указываются также показатели «степени принадлежности» членов категорий категориям — с какой частотой данный член категории был упомянут при перечислении, как часто он был упомянут первым, иногда — каков его средний порядковый номер в последовательности упоминаний. Эти нормы основываются на ответах достаточно большого числа испытуемых, перечисливших члены некоторого числа категорий. Так, нормы У. Баттига и У. Монтегю (Battig, Montague, 1969) для американского варианта английского языка основываются на ответах 442 испытуемых, перечисливших члены 56 категорий. Более современные нормы для американского английского (van Overschelde et al., 2004) основываются на ответах еще большего числа испытуемых (от 633 до 710 в зависимости от категории); перечислялись члены 70 категорий. Нами (Григорьев, 2004) были опубликованы категориальные нормы для русского языка (для 23 категорий); нормы для 20 категорий основывались на ответах 113 испытуемых, для 3 категорий — на ответах 83 испытуемых. Категориальные нормы для разных языков собирались и другими исследователями. В этих нормах зачастую приводятся данные о среднем количестве перечисленных членов каждой

категории. Эти средние широко варьируют. Например, в нашем исследовании (Там же) испытуемые перечислили в течение 30 сек., записывая свои ответы в разданных им тетрадях, в среднем 9.73 термина родства и всего 5.09 инструмента плотника. Будем называть данный показатель, отражающий зависящую от категории скорость извлечения релевантного материала из семантической памяти, продуктивностью категории. Различия категорий в продуктивности нельзя объяснить тем, что размеры категорий различаются: корреляция между продуктивностью и количеством упомянутых всеми испытуемыми членов категории в наших данных была практически нулевой (Там же, с. 74)¹. Различия лексико-семантических категорий в продуктивности обнаруживаются межъязыковую согласованность: корреляция полученных нами значений продуктивности 22 категорий с соответствующими значениями этих категорий, полученными У. Баттигом и У. Монтею, составила 0.743 (Там же).

Есть данные о различии мозговых механизмов перечисления членов разных категорий: в одном нейровизуальном исследовании было показано, что при перечислении животных и инструментов активизируются частично не совпадающие зоны мозга (Vitali et al., 2005).

Чем можно объяснить различие лексико-семантических категорий в продуктивности? В исследованиях, использовавших задачу перечисления членов категории, было замечено, что члены категории генерируются группами с небольшими временнымми интервалами между словами в группе и большими интервалами между группами; при перечислении членов лексико-семантических категорий (выполнении задачи «семантическая беглость») слова из одной группы имеют тенденцию быть семантически связанными (см.: Troyer et al., 1997). Если лексико-семантические категории различаются по степени своей семантической дифференцированности, т.е. по тому, насколько четко в них выделяются подкатегории (кластеры), содержащие семантически связанные слова, то следует ожидать и различий между ними в продуктивности: в четко дифференциированной категории искать придется в основном кластеры; переходы от слова к слову внутри кластера осуществляются без особых поисковых усилий, в то время как в нечетко дифференциированной категории объектом поиска становится большое число слов. Таким образом, мы предполагаем, что степень кластеризации лексико-семантической категории является фактором ее продуктивности.

Мнение, что одни категории могут содержать больше легко идентифицируемых группировок, чем другие, высказывалось в литературе (Diaz et al., 2004). Однако действительно ли это так? Ситуация проясняется, если для нескольких категорий будут выявлены содержащиеся в них и влияющие на ответы испытуемых кластеры, на основании чего категории можно будет сопоставлять по степени кластеризации. Для ряда категорий такие кластеры были выявлены в нашем предыдущем исследовании (Григорьев, 2004) по

¹ Мы не будем соотносить здесь данный факт с существующими представлениями о соотношении продуктивности и размера категории.

ответам испытуемых, участвовавших в сборе данных по категориальным нормам русского языка, с помощью специальной процедуры. Эта процедура будет вкратце описана ниже. По количеству кластеров в категории можно оценить степень ее кластеризации и получить ответ на вопрос, действительно ли категории различаются в этом отношении. Следующим же шагом может стать определение связи степени кластеризации категории с ее продуктивностью.

Нельзя сказать, что связь степени кластеризации перечисляемых членов лексико-семантической категории и продуктивности перечисления не рассматривалась. Так, в работе (Robert et al., 1997) сообщается о существовании такой связи. В этой работе единицами наблюдения служили испытуемые, перечислявшие члены одной лексико-семантической категории «животное». Эти испытуемые различались по количеству семантических кластеров в своих ответах, а семантические кластеры идентифицировались экспертами². Исследования, в которых единицами наблюдения были бы категории, нам неизвестны. Встречаются лишь предположения на этот счет (Diaz et al., 2004). Настоящая работа представляет собой попытку восполнить этот пробел.

Таким образом, задачей настоящей работы является проверка гипотезы, согласно которой продуктивность категории, т.е. среднее количество перечисляемых членов категории, отражающее скорость извлечения релевантного материала из семантической памяти, связано со степенью кластеризации категории. Наличие данной связи свидетельствовало бы о том, что семантическая дифференциированность категории является фактором ее продуктивности: члены категорий, которые содержат больше легко идентифицируемых группировок, будут перечисляться быстрее, т.е. испытуемые будут называть большее число членов таких категорий за единицу времени.

Метод

Материалом исследования послужили данные о продуктивности и кластеризации категорий (Григорьев, 2004). В нашей работе 2004 г., как уже указывалось, представлены нормы для 23 категорий. Из этих 23 категорий 22 были взяты из набора Баттига и Монтею, а одна («ягода») была добавлена. Для нее не была рассчитана продуктивность, поэтому в настоящем исследовании данные по ней не применялись. Используются, таким образом, данные по 22 категориям.

Оценка кластеризации категорий

Для определения того, в какие кластеры испытуемые объединяют члены категорий, нами был разработан специальный алгоритм кластерного анализа.

² К сожалению, данные авторы приводят корреляции между числом генерированных слов и числом кластеров в последовательности этих слов. Очевидно, что при постоянной степени кластеризации чем больше генерируется слов, тем больше в них будет кластеров. Если бы эти авторы представили корреляцию числа генерированных слов с отношением числа кластеров к числу генерированных слов, можно было бы делать выводы о роли кластеризации.

Он состоял в следующем. (1) В качестве исходного множества кластеров берется множество упомянутых всеми испытуемыми членов категории. (2) В качестве меры парной близости между кластерами берется отклонение наблюдаемой частоты их соседства от частоты их соседства, которую следует ожидать при случайной генерации членов категории (ожидаемой частоты), в единицах стандартного отклонения этой ожидаемой частоты. (3) Выбирается пара наиболее близких кластеров, выбранные кластеры объединяются, в дальнейшем рассматриваются в ответах испытуемых как один кластер, этот кластер включается во множество кластеров. (4) Возврат к (2). Данная процедура не является ультраметрической: кластеры могут непосредственно входить в более чем один кластер вышестоящего уровня.

Пояснения требует расчет парной близости между кластерами. Он основан на следующих соображениях. В последовательностях генерируемых испытуемыми членов категории два кластера могут либо соседствовать, либо нет. Легко показать, что вероятность случайного соседства двух кластеров, т.е. того, что два кластера случайным образом окажутся в последовательности рядом друг с другом, равна $2/n$ (n – число кластеров в последовательности), а вероятность того, что они не будут соседствовать, соответственно $1 - 2/n$. Во множестве испытаний с двумя исходами ожидаемая частота реализации одного из исходов равна сумме вероятностей реализации этого исхода в отдельных испытаниях, а ожидаемая дисперсия этой частоты – сумме произведений вероятностей двух исходов (см.: Edwards, 1964, р. 50–51). Таким образом, ожидаемая частота соседства двух кластеров во множестве последовательностей, где оба они присутствуют, равна сумме $2/n_i$ (n_i – число кластеров в i -й последовательности), а ожидаемая дисперсия этой частоты – сумме произведений $2/n_i \times (1 - 2/n_i)$. В качестве меры парной близости используется разность наблюдаемой и ожидаемой частоты соседствования двух кластеров, деленная на корень квадратный из ожидаемой дисперсии (подробнее о проблеме парной близости см.: Григорьев, 2004, с. 87).

Подчеркнем, что кластеры выделялись на основании ответов испытуемых, а не априорных соображений и мнений экспертов.

В нашем исследовании 2004 г. для каждой категории представлены 11 образовавшихся первыми объединений кластеров. В таблице 1 приводятся в качестве примеров объединения для двух категорий: «четвероногое животное» и «инструмент плотника» (по: Там же, с. 96, 108).

В качестве показателя степени кластеризации категории в настоящей работе взято количество объединений с парной близостью больше трех (т.е. с достаточно низкой вероятностью случайного соседства) среди 11 представленных кластеров. Если два кластера объединяются в один с парной близостью больше трех, считается, что они соседствуют в ответах испытуемых неслучайно, достаточно большое число испытуемых назвали соответствующие члены категории один за другим. Другими словами, мы имеем дело с устойчивым объединением членов категории. Количество таких устойчивых объединений и берется как показатель степени кластеризации категории.

Таблица 1
Примеры объединений кластеров для двух категорий

Номер объединения	Состав объединения	Близость	Состав объединения	Близость
	<i>Четвероногое животное</i>			<i>Инструмент плотника</i>
1	собака, кошка	13.08	молоток, долото	2.43
2	тигр, лев	5.78	пила, станок	2.21
3	собака, кот	5.27	гвоздь, гвоздодер	2.00
4	лошадь, корова	4.80	пила, молоток, долото	1.95
5	кошка, лошадь	3.92	рубанок, шайба	1.87
6	волк, лиса	3.89	рубанок, ключ	1.87
7	тигр, слон	3.87	долото, напильник	1.87
8	волк, заяц	3.59	пила, напильник	2.00
9	заяц, волк, лиса	3.43	рубанок, наждак	1.86
10	лев, зебра	3.42	рубанок, долото, напильник	1.73
11	мышь, крыса	3.31	пила, болгарка	1.73

Результаты и обсуждение

Средние количества перечисленных членов категории и количества объединений с парной близостью больше трех для 22 категорий представлены в таблице 2.

Как можно видеть, степень кластеризации значительно варьирует. В случае двух категорий («драгоценный камень», «инструмент плотника») в ответах испытуемых вообще не обнаруживается устойчивых группировок, в то время как в случае четырех других («четвероногое животное», «предмет одежды», «родственник», «птица») устойчивых группировок больше 10. Как и предполагали некоторые исследователи (Diaz et al., 2004), одни категории содержат больше легко идентифицируемых группировок, чем другие. По сути дела, продемонстрирована структурная неоднородность лексико-семантических категорий. Существование структурной неоднородности лексико-семантических категорий, рассматриваемых как языковые образования, а не с точки зрения их представленности в семантической памяти, известно лингвистам (см.: Филлмор, 1983). Однако в нашем исследовании эта неоднородность выявлена по ответам испытуемых, что дает право рассматривать ее в психологическом плане, полагая, что она характеризует репрезентации категорий в семантической памяти. Таким образом, помимо того, что различные лексико-семантические категории кодируются, по-видимому, разнородными признаками, их репрезентации, можно полагать, также являются структурно неоднородными.

Таблица 2

Средние количества перечисленных членов категории и количества объединений с парной близостью больше трех для 22 категорий

Категория	Количество объединений с парной близостью больше трех	Среднее количество перечисленных членов категории
Драгоценный камень	0	5.45
Четвероногое животное	11	8.26
Вид ткани	0	6.00
Цвет	8	8.48
Предмет мебели	9	8.05
Овощ	6	6.99
Предмет одежды	11	8.66
Вид транспортного средства	7	7.08
Кухонная посуда	10	7.53
Вид спорта	7	6.78
Наука	3	6.10
Рыба	4	7.43
Родственник	11	9.73
Часть здания	5	6.25
Цветок	2	7.28
Фрукт	7	7.16
Дерево	6	8.56
Птица	11	8.45
Профессия	2	6.81
Металл	2	5.90
Инструмент плотника	0	5.09
Музикальный инструмент	5	7.37

Коэффициент корреляции, равный произведению моментов (Пирсона), для переменных в таблице 2 составляет 0.838; $p < 0.001$; коэффициент ранговой корреляции (Спирмена) составляет 0.817; $p < 0.001$. Таким образом, гипотеза о связи кластеризации категорий и скорости перечисления ее членов подтвердилась. Это согласуется с предположением, что степень кластеризации лексико-семантической категории (и ее презентации) является фактором ее продуктивности. Возможная структурная неоднородность лексико-семантических категорий оказалась подходящим объяснением их различий в продуктивности.

В заключение отметим, что структурная неоднородность лексико-семантических категорий не может быть сведена к различию в степени кластеризации. Действительно, простой количественный показатель не дает указаний на то, что могут существовать разные типы структур таких объектов. Так, лингвисты (Там же) выделяют разные типы структур семантического поля (по сути дела, лексико-семантической категории): партономия, парадигма и т.д. Тип структуры лексико-семантической категории также может быть фактором ее продуктивности. Так, кажется вероятным, что высокая продуктивность категории «родственник» является в значительной степени следствием того, что данная категория представляет собой прототип парадигмы. Действительно, то, что одна оппозиция терминов родства (по полу родственника) регулярно повторяется на разных градациях других оппозиций (по поколению, родственной близости), значительно облегчает задачу испытуемого.

Добавим, что показатель степени кластеризации категории может подлежать разным истолкованиям при разных типах ее структуры. Например, в случае парадигмы высокая кластеризация обусловлена большим количеством парных объединений, основывающихся на небольшом числе одних и тех же связей. При другом же типе структуры — таксономии — за высокой кластеризацией может стоять меньшее число не обязательно парных объединений, основывающихся на разнообразных связях.

Литература

- Алфимова, М. В. (2010). Семантическая вербальная беглость: нормативные данные и особенности выполнения задания больными шизофренией. *Социальная и клиническая психиатрия*, 20(3), 20–25.
- Григорьев, А. А. (2004). *Репрезентация лексических категорий в сознании носителя языка*. М.: Институт языкознания РАН.
- Дроздова, К. А., Рупчев, Г. Е., Семенова, Н. Д. (2015). Нарушение вербальной беглости у больных шизофренией. *Социальная и клиническая психиатрия*, 25(4), 9–19.
- Филлмор, Ч. Дж. (1983). Об организации семантической информации в словаре. В кн. *Новое в зарубежной лингвистике* (вып. 14, с. 23–60). М.: Прогресс.

Ссылки на зарубежные источники см. в разделе References после англоязычного блока.

Григорьев Андрей Александрович — главный научный сотрудник, ФГБУН «Институт психологии РАН», доктор филологических наук, доцент.
Сфера научных интересов: интеллект, индивидуальные различия, психолингвистика.
Контакты: andrey4002775@yandex.ru

The Relationship between the Degree of Clusterization of Lexical-Semantic Category and the Speed of Recitation of Its Objects

A.A. Grigoriev^a

^aInstitute of Psychology of Russian Academy of Sciences, 13 build. 1, Yaroslavskaya str., Moscow, 129366, Russian Federation

Abstract

In the task of reciting objects belonging to lexical-semantic categories (category members), participants recite the members of some categories faster than the members of another. Results obtained in the studies on this task suggest that the degree of clusterization of categories' members may be the factor, which causes this difference between categories: categories, which members are grouped better have an advantage. To verify this suggestion, categorical norms for Russian language, which contain data about the average numbers of categories' members recited by participants per unit of time, and about the clusters in categories, which affect participants' responses, were used. Data on 22 categories (Precious Stone, 4-footed Animal etc.) were analyzed; for each of them the number of clusters with pair proximity above 3 was calculated; this number was used as a measure of degree of clusterization. It turned out that the categories significantly vary with respect to this measure: while for two categories (Precious Stone, Carpenter's Tool) it is zero, for four others (4-footed Animal, An Article of Clothing, A Relative, A Bird) it is above 10. The correlation between degree of clusterization and the mean number of category members generated was .838; $p < .001$. This result is consistent with the suggestion about the role of degree of clusterization of the members of lexical-semantic categories as a factor of the speed of recitation of objects belonging to it. However, other differences in structure of lexical-semantic categories may be such factors; in particular, a type of this structure may play a part.

Keywords: semantic memory, representation, semantic category, category norms, clusterization.

References

- Alfimova, M. V. (2010). Semantic verbal fluency: normative data and task performance in a schizophrenic sample. *Sotsial'naya i Klinicheskaya Psichiatriya*, 20(3), 20–25. (in Russian)
- Battig, W. F., & Montague, W. E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut norms. *Journal of Experimental Psychology*, 80, 1–46.
- Bokat, C. E., & Goldberg, T. E. (2003). Letter and category fluency in schizophrenic patients: a meta-analysis. *Schizophrenia Research*, 64, 73–78.
- Diaz M., Sailor, K., Cheung, D., & Kuslansky, G. (2004). Category size effects in semantic and letter fluency in Alzheimer's patients. *Brain and Language*, 89, 108–114.
- Drozdova, K. A., Rupchev, G. E., & Semenova, N. D. (2015). Verbal fluency deficits in schizophrenia. *Sotsial'naya i Klinicheskaya Psichiatriya*, 25(4), 9–19. (in Russian)
- Edwards, A. L. (1964). *Expected values of discrete random variable and elementary statistics*. New York: John Wiley and Sons.

- Fillmore, Ch. J. (1983). Ob organizatsii semanticheskoi informatsii v slovare [On the organization of semantic information in the dictionary]. In *Novoe v zarubezhnoi lingvistike* [New in the foreign linguistics] (Iss. 14, pp. 23–60). Moscow: Progress. (in Russian)
- Grigoriev, A. A. (2004). *Reprezentatsiya leksicheskikh kategorii v soznanii nositelya yazyka* [Representation of lexical categories in the native speaker's consciousness]. Moscow: Institute of Linguistics of the RAS. (in Russian)
- Pilgrim, L. K., Moss, H. E., & Tyler, L. K. (2005). Semantic processing of living and nonliving concepts across the cerebral hemispheres. *Brain and Language*, 94, 86–93.
- Robert, P. H., Migneco, V., Marmod, D., Chaix, I., Thauby, S., Benoit, M., ... Darcourt, G. (1997). Verbal fluency in schizophrenia: the role of semantic clustering in category instance generation. *European Psychiatry*, 12, 124–129.
- Troyer, A. K., Moscovitch, M., & Winocur, G. (1997). Clustering and switching as two components of verbal fluency: evidence from younger and older healthy adults. *Neuropsychology*, 11(1), 138–146.
- Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language*, 50, 289–335.
- Vitali, P., Abutalebi, J., Tettamanti, M., Rowe, J., Scifo, P., Fazio, F., ... Perani, D. (2005). Generating animal and tool names: An fMRI study of effective connectivity. *Brain and Language*, 93, 32–45.
- Warrington, E. K., & Shallice, T. (1984). Category specific semantic impairments. *Brain*, 107, 829–854.

Andrei A. Grigoriev — chief research fellow, Institute of Psychology of Russian Academy of Sciences, D.Sc., associate professor.
Research area: intelligence, individual differences, psycholinguistics.
E-mail: andrey4002775@yandex.ru