Глава 13.
К проблеме обработки экспериментальных данных*

Данная глава содержит обзор процедур, которые были использованы при обработке результатов экспериментов, представленных в настоящей монографии. В главе рассматриваются особенности практического использования методики вынесения суждений, а также необходимые шаги предварительной систематизации полученных данных и статистической обработки. Глава также содержит краткий обзор использованных статистических критериев. Авторы и редакторы выражают надежду, что рекомендации из данного обзора будут полезны для читателей, которые только начинают свой путь в сборе экспериментальных данных для решения собственных исследовательских задач.

13.1. Введение

Исследования островных ограничений относятся к тому типу феноменов, для изучения которых необходимо использование экспериментальных данных. Это связано прежде всего с низкой частотностью изучаемых конструкций и специфичностью грамматической структуры, ввиду чего становится невозможным поиск данных в корпусе. Кроме того, сформулированные гипотезы касаются очень частных аспектов языковой структуры, манипулирование которыми возможно только в рамках лингвистического эксперимента.

В качестве основного метода эмпирической проверки формально-грамматических принципов использовалась совокупность процедур, разработанных в рамках парадигмы экспериментального синтаксиса ([Schütze 1996; Cowart 1997; Featherston 2007] и др.). В основе этих процедур лежит факторный дизайн, посредством которого предполагаемое ограничение раскладывается («факторизуется») путем изоляции потенциальных источников понижения приемлемости. Далее обнаруженное взаимодействие

* Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 20-312-90004.

в результатах эксперимента может быть интерпретировано как следствие наличия грамматического ограничения. Данная парадигма позволяет осуществлять самые разные разложения на факторы с учетом как грамматических, так и потенциально экстраграмматических факторов (ср. исследование островных ограничений в рамках редукционистского подхода [Sprouse et al. 2012]) и их сочетаний (в том числе манипулируемый фактор может варьироваться между разными экспериментами с одним и тем же дизайном).

13.2. Вопросы, связанные с обработкой оценок приемлемости

Основной экспериментальной методикой, которая использовалась в проекте «Русские острова в свете экспериментальных данных», стала методика извлечения суждений. В главе 6, посвященной асимметрии выдвижения аргументов и адъюнктов, и в главе 10 о природе эффектов превосходства также использовалась методика чтения с саморегуляцией скорости. Суждения о грамматичности в основном выносились по шкале Ликерта от 1 до 7. Единственным исключение стало исследование интрузивных местоимений (глава 11), в котором также использовался метод выбора между альтернативами (*forced choice*).

Оценка по шкале Ликерта подразумевает, что респондент выносит оценку по шкале, крайние значения которой соответствуют приемлемому или неприемлемому предложению. Традиционный выбор шкал от 1 до 5 или от 1 до 7 обусловлен тем, что, с одной стороны, на них можно обозначить середину, с другой стороны, количество точек на шкале считается достаточным для обнаружения различий между минимальными парами языковых выражений. Существенным достоинством шкалы Ликерта является тот факт, что эта методика позволяет численно охарактеризовать величину различия между условиями [Schütze, Sprouse 2014]. Кроме того, обнаруживается, что данный метод является более мощным в сравнении с другими числовыми методами извлечения суждений [Sprouse et al. 2018] и одновременно характеризуется существенной корреляцией результатов с результатами применения как нечисловых методов извлечения суждений, так и интроспекции [Langsford et al. 2018].

Тем не менее, имеет смысл упомянуть один значимый недостаток этого метода, а именно: исследователи не могут однозначно утверждать, каким образом происходит вынесение оценки по шкале, то есть какие

процессы вовлечены в использование шкалы респондентом. Этот недостаток неизбежен ввиду проблемы черного ящика, то есть невозможности наблюдать когнитивные процессы.

Одно из следствий названного недостатка состоит в том, что мы не знаем, как устроены интервалы на шкале приемлемости. Например, неясно, оценивают ли респонденты различие между 1 и 2 точно так же, как различие между 4 и 5. От ответа на этот вопрос зависит, к какой шкале измерений будет относиться получаемая переменная (оценка). Если предположить, что интервалы между точками на шкале равны, тогда речь будет идти о шкале интервалов. Если же интервалы отличаются, шкала измерений будет порядковая, то есть предполагающая ранжирование значений переменной. Другими словами, мы наблюдаем противопоставление качественных дискретных шкал и количественных непрерывных.

Названный факт вызывает значительные дискуссии о том, какие статистические методы стоит использовать при обработке результатов применения шкалы Ликерта (см. историю вопроса и обзор аргументов в [Harpe 2015])[62]. По этому вопросу можно выделить несколько точек зрения. Так, в рамках подхода, представленного в резонансной заметке [Jamieson 2004], предлагается считать шкалу Ликерта порядковой шкалой. Это означает, что к данным возможно применять только непараметрические критерии, указывать выборочное среднее и стандартное отклонение некорректно.

Однако достаточное количество авторов критикуют такой радикальный взгляд на проблему. Так, в соответствии с центральной предельной теоремой предлагается считать, что распределение средних или остатков регрессии (*residuals*) для выборок, не относящихся к малым (от 5 до 10 респондентов), будет близким к нормальному, чего достаточно для применения как t-критерия Стьюдента, так и дисперсионного анализа [Norman 2010]. В комментарии [Carifio, Perla 2008] также отмечается, что t-критерий Стьюдента, например, очень устойчив к отклонению от нормальности, в пользу чего свидетельствуют множественные исследования, основанные на методе симуляций выборок. Авторы приходят к тому, что

[62] В данной главе мы не претендуем на полноту обзора проблематики использования шкалы Ликерта. Упомянутая дискуссия продолжается с того времени, когда шкала была предложена [Likert 1932]. Заметим также, что в области лингвистики подобный вопрос отдельно не рассматривался, основные исследования проводились в области психологии, а также в области социальных наук и образования.

для анализа совокупности или суммы оценок на шкале Ликерта вполне допустимо использовать параметрические методы. Аналогичные исследования с симуляциями проводились и для дисперсионного анализа. Так, в исследовании [Schmider et al. 2010] показано, что результаты применения дисперсионного анализа к выборкам из генеральных совокупностей, имеющих различное распределение, являются устойчивыми с точки зрения вероятностей ошибок первого и второго рода. Более того, используя метод регрессионного анализа, авторы демонстрируют, что фактор типа распределения данных, в отличие от фактора размера эффекта, является незначимым. В работе [Mircioiu, Atkinson 2017] две группы статистических подходов, параметрические и непараметрические, применяются к данным реального исследования. Показано, что методы дают идентичные результаты касательно значимости ожидаемых различий.

Важно выделить одно ключевое соображение: поскольку оценки на шкале Ликерта дискретны, а также ограничены сверху и снизу, в целом говорить о каком-либо распределении этих оценок некорректно. Тем не менее, выделяется ряд практических приемов, которые используются для того, чтобы исследователь с большей уверенностью мог использовать параметрические критерии. Так, работа [Wu, Leung 2017] является одним из исследований, в которых обосновывается следующий вывод: большее количество точек на шкале Ликерта способствует снижению влияния фактора дискретности шкалы, а следовательно, ведет к большей близости к нормальному распределению.

В отношении применения шкалы Ликерта в лингвистических исследованиях можно выделить два практических приема, косвенно имеющих отношение к решению упомянутой проблемы. Во-первых, одним из нормативов экспериментального синтаксиса уже стала нормализация исходных данных [Schütze, Sprouse 2014]. Нормализация используется в первую очередь для решения проблемы искажения шкалы (*scale bias*). Существует распространенное мнение, что респонденты по-разному используют шкалы оценок. Индивидуальные стратегии могут проявляться в постоянном использовании тех или иных областей шкалы, использовании только полярных значений, или, напротив, преднамеренном исключении крайних значений. Нормализация решает две задачи: а) оценки респондента центрируются относительно средней оценки по всем ответам респондента; б) в качестве меры расстояния текущей оценки от среднего на шкале используется стандартное отклонение, посчитанное по всем ответам респондента. Формально **z-оценка** (**нормализованная оценка** или ***z-score***) определяется следующим образом:

$$Z_{ij} = \frac{(X_{ij} - \overline{X}_i)}{\sigma_i}$$

где: X_{ij} — j-ая оценка i-ого респондента;

$\overline{X}_i$ — выборочное среднее i-ого респондента по всем экспериментальным условиям;

σ_i — стандартное отклонение i-ого респондента по всем экспериментальным условиям;

Z_{ij} — нормализованная j-ая оценка i-ого респондента.

Заметим, однако, следующую вещь. Нормализованные оценки можно принять за значения непрерывной переменной (ввиду различия в величинах стандартного отклонения для разных респондентов теряется привычный вид шкалы с точками, фиксированными на равном расстоянии друг от друга). Тем не менее, нормализация является именно линейным преобразованием (в отличие от, например, перевода данных из ранговой шкалы в количественную с применением аппарата теории нечетких множеств [Каган, Маруцак 2014]), то есть применение нормализации к исходным оценкам никак не влияет на шкалу измерений. Более того, как было уже сказано выше, в соответствии с радикальным взглядом на природу оценок по шкале Ликерта подсчет выборочного среднего и стандартного отклонения некорректен, а следовательно, в рамках такого подхода некорректной считается и нормализация. Можно сделать вывод, что в большинстве лингвистических экспериментальных исследований полученные данные по умолчанию рассматриваются с допущением насчет их интервальной природы.

Также в исследования по экспериментальному синтаксису часто включаются филлеры различного уровня приемлемости. В обучающих материалах [Sprouse 2018] этот шаг обосновывается следующим образом. По мнению Дж. Спрауза, респонденты следят за тем, как часто они используют разные значения на шкале. Если какие-то значения не используются длительное время, респонденты могут начать использовать эти значения даже тогда, когда эти значения неуместны. Данную проблему могут решить разнообразные по приемлемости филлеры: они помогают вынуждать респондента использовать максимальное количество значений шкалы. Таким образом, обеспечивается равномерное использование шкальных значений, что в свою очередь также способствует меньшей смещенности получаемых данных.

Аргумент Дж. Спрауза не был подтвержден экспериментально. Важно сказать, что в настоящей монографии исследователи активно используют различные филлеры, но делается это по другим причинам. В соответствии с принятыми в экспериментальном синтаксисе процедурами обработки данных, для каждого из проведенных экспериментов проводилась нормализация индивидуальных оценок респондентов. На наш взгляд, нормализация имеет существенное ограничение, проявляющееся в том, что уже невозможно наблюдать размерность шкалы и ее крайние значения. Заметим, что все респонденты получали осмысленную инструкцию насчет значения отдельных точек на шкале приемлемости. В связи с этим было бы полезно понимать, к какой области шкалы относятся те или иные конструкции (например, являются они скорее приемлемыми или скорее неприемлемыми). Грамматичные и неграмматичные[63] филлеры позволяют обозначить границы шкалы после нормализации. Это не только облегчает визуальное восприятие экспериментальных результатов (см. рис. 13.1), но и дает возможность проверить, есть ли значимые отличия неприемлемых экспериментальных предложений от конструкций, в которых нарушены известные грамматические запреты. Другими словами, за счет введения полярных значений осуществляется локализация оценок для отдельных условий на шкале приемлемости.

Итак, в настоящей монографии мы принимаем позицию, согласно которой говорить о характеристике распределения оценок в принципе некорректно. Тем не менее, из практических соображений, подтвержденных систематическими исследованиями с симуляциями выборок, а также в соответствии с принятыми нормами проведения синтаксических экспериментов мы проводим нормализацию данных и далее анализируем результаты, исходя из предположения о том, что нарушения требований к использованию параметрических критериев являются несущественными.

---

[63] Мы не случайно используем здесь понятие грамматичности. В то время как грамматически правильными считаются цепочки слов, которые могут быть порождены некоторой грамматикой, приемлемыми называются цепочки слов, которые носители языка признают в качестве правильных предложений своего языка. Грамматичность представляет собой лишь один из факторов, определяющих приемлемость. Строго говоря, респонденты не могут оценить грамматичность, поскольку грамматика есть ментальный конструкт. Тем не менее, в данном случае мы хотим охарактеризовать филлеры именно с точки зрения грамматичности, то есть того, могут ли они быть порождены в рамках грамматики. Ожидается, что неграмматичные филлеры, то есть предложения, нарушающие критические грамматические ограничения, получат наименьшую оценку приемлемости из доступных.

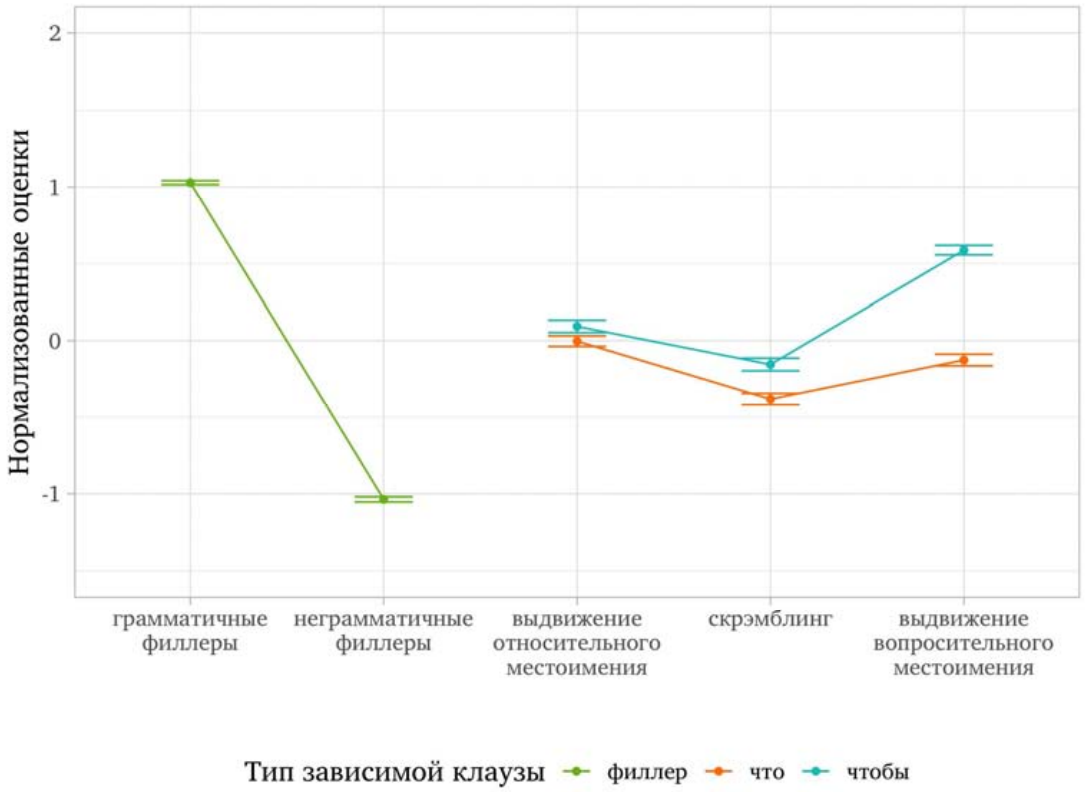

Рисунок 13.1. Результаты экспериментального исследования островных свойств придаточных изъяснительных с союзом *что* (глава 4)

13.3. Систематизация и предварительная обработка экспериментальных данных

Прежде чем прибегать к конкретным статистическим методам, необходимо подготовить данные к проведению статистического анализа. Подготовку можно разделить на два шага: форматирование и преобразование исходных результатов.

Перечислим основные действия, которые авторы исследований совершали с целью систематизации ответов респондентов. В первую очередь все условия были закодированы, были анонимизированы имена респондентов, отдельно проанализированы личные данные участников исследований (такие как возраст, город проживания, наличие лингвистического образования и т.п.). Важным шагом при систематизации является перевод данных из широкого табличного формата, в котором строки соответствуют

отдельным респондентам, а столбцы — графам ответов, в длинный, в котором строка представляет собой наблюдение, а столбцы кодируют характеристики этого наблюдения (к характеристикам относится как респондент, который вынес оценку, так и сама оценка и уровни независимых переменных, задающие конкретное условие)[64]. Именно длинный формат будет наиболее удобным (а в каких-то случаях единственно возможным) для обработки экспериментальных данных в среде R [R Core Team 2017].

Преобразование данных включает в себя две возможные операции. В первую очередь исключаются аномальные ответы. Далее следуют преобразования, диктуемые характеристиками данных (например, нормализация в случае оценок приемлемости). Отдельный вопрос состоит в том, что следует считать аномалией. В текущем проекте авторы использовали одновременно несколько критериев отбора респондентов.

Наиболее прямой способ отбора респондентов подразумевает сравнение ответов с заданным заранее абсолютным интервалом допустимых значений. Так, в проведенных исследованиях исключались респонденты, время ответа которых было слишком малым или, напротив, увеличенным. Другой возможный вариант отбора может опираться на относительные интервалы допустимых значений, заданные в зависимости от среднего значения по выборке или по ответам одного респондента. Поскольку в проведенных исследованиях респонденты не совершали действий, которые существенно зависели бы от индивидуальных качеств, данный вариант не использовался. Тем не менее, он может оказаться действенным в отборе респондентов по такому параметру, как время реакции.

Наконец, резко отклоняющиеся значения можно определить по ряду контрольных параметров. В проекте в качестве таких параметров использовались контрольные вопросы на понимание содержания к грамматичным филлерам, а также оценки грамматичных и неграмматичных филлеров: не учитывались ответы респондентов, которые часто ошибались или ставили крайние положительные или крайние отрицательные оценки неграмматичным и грамматичным филлерам соответственно. В ряде работ также использовался описанный Дж. Спраузом метод подсчета суммы квадратов разниц реальных и ожидаемых оценок для филлеров с последующей процедурой винсоризации [Sprouse 2018].

[64] Основной программой для проведения экспериментов стала платформа Ibex farm [Drummond 2013], так что в представленных исследованиях этот шаг был пропущен ввиду того, что результаты сразу записываются в длинном формате.

13.4. Визуализация данных

Прежде чем применять конкретные статистические методы, необходимо составить количественное описание выборки посредством основных статистических показателей. Дескриптивная статистика позволит сделать предварительные выводы на основе имеющихся данных и принять стратегические решения для статистического анализа. Не менее важна на данном этапе визуализация данных. В исследованиях, представленных в настоящей монографии, для разных целей использовались различные типы визуализации.

Во-первых, для демонстрации средних оценок и стандартных отклонений для различных условий использовались диаграммы размаха. Далее, для визуализации взаимодействия факторов использовались графики взаимодействия (*interaction plot*), на которых указывались средние и стандартные ошибки для всех условий. Наконец, в некоторых исследованиях возникало подозрение о существовании бимодального распределения для определенных условий. Это могло означать, что в выборку попали респонденты из разных генеральных совокупностей, то есть обладающие различной грамматикой. Для демонстрации таких ситуаций использовались гистограммы оценок приемлемости. Пример анализа бимодального распределения и разделения респондентов на группы можно увидеть в главе 12, посвященной исследованию эффекта комплементайзер-след.

13.5. Статистическая обработка оценок приемлемости

Для обработки материала использовались статистические методы, принятые в настоящий момент в качестве стандартных для анализа экспериментальных данных. Сначала рассмотрим методы, применявшиеся для анализа оценок приемлемости. Наименее трудоемким и самым распространенным методом является **многофакторный дисперсионный анализ** (с единичными или с повторными измерениями), который основан на сравнении внутригрупповой изменчивости с общей изменчивостью. Дисперсионный анализ помогает установить, значимо ли влияние отдельных факторов (независимых переменных) и их взаимодействия на наблюдения (зависимую переменную). Для последующих множественных попарных сравнений отдельных условий между собой использовался **критерий Тьюки**.

Несколько более трудоемким, но универсальным статистическим методом являются **смешанные линейные модели**, которые позволяют одновременно учитывать вариативность за счет потенциально неограниченного числа так называемых «случайных» эффектов, которые специально не контролируются в эксперименте. К таким эффектам относятся, в частности, влияние конкретного экспериментального предложения и влияние конкретного участника эксперимента. Кроме того, смешанные линейные модели успешно применяются к наборам данных, в которых по каким-то причинам отсутствует некоторое количество точечных измерений[65].

Далее, в некоторых случаях возникала необходимость в сравнении средних для отдельных условий вне многофакторной модели (например, в главе 9 нужно было провести сопоставление между парами предложений с разным порядком слов и одинаковым статусом по отношению к дискурсивной связанности). Для таких случаев использовался **t-критерий Стьюдента** или непараметрический **критерий Вилкоксона** с поправкой Бонферрони на множественную проверку гипотез.

В главе 11 в эксперименте также использовался метод выбора между альтернативами. Для анализа данных нечисловых шкал в первую очередь использовался **критерий χ^2**, который для полученной таблицы сопряженности показывает, связаны ли переменные. Для того чтобы проверить значимость различий для двух альтернатив, использовался **критерий знаков** (реализованный в R как частный случай **биномиального теста**). Для сопоставления двух альтернатив также можно воспользоваться **коэффициентом (фактором) Байеса**, который дает количественную оценку адекватности данным одной модели по сравнению с другой вне зависимости от истинности этих моделей.

13.6. Заключение

Несмотря на то, что методы экспериментального синтаксиса становятся все более востребованными для решения отдельных исследовательских задач, до сих пор остается ряд нерешенных вопросов как по процедуре проведения эксперимента, так и по особенностям обработки результатов. В данной главе мы описали и обосновали выработанный в проекте взгляд на обработку экспериментальных данных, прежде всего оценок приемлемости. Мы подробно рассмотрели проблему отнесения

[65] О применении смешанных линейных моделей в анализе результатов лингвистических экспериментов см. подробнее [Bross 2019; Winter 2013].

оценок по шкале Ликерта к порядковой или интервальной шкале измерений, раскрыли физический смысл нормализации оценок, аргументировали использование грамматичных и неграмматичных филлеров и сопоставление с ними целевых условий эксперимента. Затем мы описали порядок систематизации и трансформации данных, привели различные типы визуализации в соответствие с целями исследователей. Наконец, мы рассмотрели использованные в монографии статистические критерии.

Мы предполагаем, что открытость исследователей в отношении решений, определяющих дизайн и процедуру эксперимента, а также конкретных способов обработки полученных результатов способствует не только формированию более точного представления о происходящем у читателя, но и укреплению в научном сообществе идеи о воспроизводимости исследований, в том числе и в традиционно гуманитарных науках.