Гращенков Павел Валерьевич

Московский государственный университет им. М. В. Ломоносова,
Россия, 119991, Москва, Ленинские горы, 1
https://orcid.org/0000-0001-9754-2452
pavel.gra@gmail.com

Студеникина Ксения Андреевна

Московский государственный университет им. М. В. Ломоносова,
Россия, 119991, Москва, Ленинские горы, 1
https://orcid.org/0000-0002-4098-7167
xeanst@gmail.com

Паско Лада Игоревна

Московский государственный университет им. М. В. Ломоносова,
Россия, 119991, Москва, Ленинские горы, 1
https://orcid.org/0000-0002-0533-809X
paskolada@yandex.ru

Ограничение сочинительного острова в лингвистической компетенции больших языковых моделей*

Для цитирования: Гращенков П. В., Студеникина К. А., Паско Л. И. Ограничение сочинительного острова в лингвистической компетенции больших языковых моделей. *Вестник Санкт-Петербургского университета. Язык и литература*. 2024, 21 (3): 668–688.
https://doi.org/10.21638/spbu09.2024.309

Синтаксическим островом называется конструкция, извлечение элемента из которой приводит к неграмматичности. Действие островных ограничений обычно демонстрируется на материале передвижений операторного типа — например, вопросительного передвижения. Восприятие извлечения из сильных синтаксических островов как неграмматичного присуще всем носителям естественного языка. В настоящее время большие языковые модели способны успешно вести диалог с пользователем на различные темы: они понимают смысл запросов и порождают текст, неотличимый от человеческого. Однако остается малоизученным вопрос о том, насколько схожи грамматические ограничения, которые накладываются на языковую способность людей и нейросетей. В том случае, если грамматика больших языковых моделей идентична человеческой, мы также ожидаем от них высокую чувствительность к нарушению языковых ограничений. Цель исследования состоит в том, чтобы сравнить языковую способность носителей языка и нейросетевых моделей на основе интерпретации острова сочинительной конструкции. Была проанализирована лингвистическая компетенция трех русскоязычных диалоговых моделей — ChatGPT, YandexGPT и GigaChat — с помощью двух тестов. Первый проверяет, способна ли модель верно ответить на вопрос, образованный с нарушением островных ограничений. Второй тест напрямую обращает-

* Работа выполнена при поддержке Программы развития МГУ, проект № 23-Ш02-10 «Языковая компетенция носителей естественного языка и нейросетевых моделей».

ся к оценке грамматичности предложения. Результаты показали, что лингвистическая компетенция больших языковых моделей отличается от человеческой. Количество семантически верных ответов и положительных оценок грамматичности оказалось достаточно большим для всех моделей. Поведение YandexGPT является наиболее логичным, тогда как модели ChatGPT и GigaChat часто были не готовы отвечать на вопросы, которые сами считали корректными. Обнаружилось, что грамматические характеристики стимульных предложений по-разному влияют на способность моделей отвечать на вопросы и оценивать их корректность: ChatGPT и GigaChat демонстрируют схожее поведение в противоположность YandexGPT. Результаты исследования ставят под сомнение утверждение о том, что поведение больших языковых моделей идентично поведению людей.

Ключевые слова: большие языковые модели, обработка естественного языка, островные ограничения, русский язык, синтаксис.

Введение

Ограничение на вынос из сочиненной структуры

Островными ограничениями в синтаксисе называется невозможность или затрудненность извлечения материала из конструкций определенного типа — островов, см. исследование [Ross 1967] и последовавшие за ним работы. Под извлечением обычно понимаются операции так называемого A′-передвижения — передвижения в неаргументные позиции, например образование вопроса и относительной клаузы, топикализация. Острова, исключающие любое выдвижение, называются сильными. К сильным относятся сочинительный остров, остров сложной именной группы, субъектный остров, обстоятельственный остров. Из слабых островов выдвижение затруднено только для части составляющих. Слабыми являются острова косвенного вопроса, отрицательный и некоторые другие (подробнее об этом см.: [Cinque 1990; Лютикова, Герасимова 2021; и др.]).

В нашем исследовании мы рассмотрим остров, образуемый составляющими с сочинительными союзами (*и*, *или*, *а* и др.)[1]. В русском языке извлечение из острова сочинительной конструкции ведет к неграмматичности предложения — остров является сильным. Так, в примерах в (1) образование вопроса невозможно, если вопросительное слово соотносится с любым из двух конъюнктов в утвердительном предложении (1b,c) или входит в состав конъюнкта (1b’,c’,b”,c”).

(1) a) В этой комнате помещаются [два дивана и компьютер].
b) *Что в этой комнате помещаются [~~что~~ и компьютер]?
c) *Что в этой комнате помещаются два дивана [и ~~что~~]?

a’) Маша хочет, чтобы [[Вася починил посудомойку], а [Лена подмела пол]].
b’) *Кто Маша хочет, чтобы [[~~кто~~ починил посудомойку], а [Лена подмела пол]]?
c’) *Кто Маша хочет, чтобы [[Вася починил посудомойку] и [~~кто~~ подмел пол]]?

[1] Согласно исследованию Дж. Росса, «[в] сочиненной структуре никакой конъюнкт не может быть перемещен, равно как и никакой элемент из состава конъюнктов не может подвергнуться передвижению» (исходная цитата: «In a coordinate structure, no conjunct may be moved, nor may any element contained in a conjunct be moved out of that conjunct» [Ross 1967: 162], перевод на русский язык выполнен авторами данной статьи).

a”) Маша хочет, чтобы [[Вася починил посудомойку], а [Лена подмела пол]].
b”) *Что Маша хочет, чтобы [[Вася починил ~~что~~], а [Лена подмела пол]]?
c”) *Что Маша хочет, чтобы [[Вася починил посудомойку], а [Лена подмела ~~что~~]? [Моргунова 2021: 37].

Есть два известных случая отклонений от ограничения на вынос из сочинительного острова. Во-первых, это так называемое «поперечное» (Across-the-board, ATB) передвижение (см.: [Williams 1978]). В этом случае один и тот же участник выносится из обеих конъюнктов:

(2) Какие книги Маша любит ~~какие книги~~, а Петя ненавидит ~~какие книги~~?

Во-вторых, сочиненные глагольные группы демонстрируют так называемое псевдосочинение [Lakoff 1986], образуя конструкции, где набор участников не одинаков и имеющийся в одном из конъюнктов участник может подвергнуться передвижению:

(3) a) Это чай, который ты ушел в кино и не допил.
b) Там был еще один юноша, который всем надоел, и его выставили [Зализняк, Падучева 1979: 305].

Для объяснения феномена островных ограничений в разных версиях генеративной грамматики привлекается разный инструментарий. В актуальной версии, минимализме, запрет на передвижение объясняется как результат взаимодействия между синтаксисом и модулями интерпретации и озвучивания. Передвижения должны быть циклическими и происходят между так называемыми фазами — структурными блоками, которые должны получать законченную интерпретацию и озвучивание. Между фазами для выдвижения могут быть доступны лишь определенные зоны в определенных конфигурациях. Передвижение может быть невозможно в силу нарушения цикличности или неблагоприятной конфигурации элементов внутри фаз (см.: [Chomsky 2004; 2013]).

Наряду с данным грамматическим подходом, предполагающим врожденность островных ограничений, существует так называемый редукционистский подход. Согласно ему, островные ограничения не связаны с конкретными типами синтаксических конструкций напрямую. Проблемы с грамматичностью и интерпретируемостью предложений, нарушающих островные ограничения, следуют из общей сложности таких структур. Для них, например, необходимо устанавливать дистантные (long-distance) отношения между вопросительным словом и его следом в исходной структуре, причем позиций для следа может быть потенциально больше одной. Редукционизм per se при этом оказывается неспособен последовательно объяснить все множество наблюдаемых в связи с островами явлений (см. [Phillips 2013a; Лютикова, Герасимова 2021: 25–26]). Авторы данной работы стоят на тех позициях, согласно которым островные ограничения носят абсолютный (т. е. грамматический) характер, факторы же структурной сложности могут усугублять неграмматичность.

Отдельный вопрос — причины существования островных ограничений в конкретных языках. Согласно грамматическому подходу, которого придерживаются сторонники генеративной грамматики, островные ограничения являются след-

ствием врожденных принципов универсальной грамматики и не могут появляться в сознании носителей в процессе усвоения языка. Противоположная точка зрения заключается в том, что островные ограничения усваиваются в процессе овладения языком в детстве, см. дискуссию в работах [Phillips 2013b; Pearl, Sprouse 2013; и др.]. В общем виде усвоение неприемлемости островов представляется при «несинтаксическом» подходе как результат «вероятностного анализа» входных данных при усвоении языка ребенком. Наблюдая, что дистантные отношения филлера и пробела через несколько клауз крайне редки, дети усваивают информацию о нежелательности их использования[2].

Исследования языковых способностей нейросетевых моделей

Вопрос усвоения или врожденности ограничений важен в контексте поставленной в нашей работе задачи, так как большие языковые модели (БЯМ) приобретают языковую компетенцию исключительно из «увиденных» ими текстов. Общий принцип работы языковых моделей, в том числе и БЯМ, поведение которых мы будем исследовать в данной работе, заключается в предсказании наиболее вероятного продолжения последовательности слов или символов. Наиболее естественным и вероятным продолжением, соответственно, будет то, которое частотнее в обучающей выборке.

В целом можно сказать, что от человека и БЯМ ожидается разная реакция на неграмматичные структуры. Человеку свойственно отказываться от интерпретации тех структур, которые представляются ему неграмматичными. Исследуя представление островных и других грамматических ограничений в БЯМ, мы не просто изучаем их «языковую интуицию», но и пытаемся понять границы их возможностей в интерпретации. Они могут быть более узкими или более широкими, чем человеческие, или могут просто не совпадать с таковыми у людей.

Некоторые проведенные предыдущими исследователями эксперименты с компьютерным моделированием имели целью имитацию усвоения синтаксических островов на основе ограниченного языкового материала (см.: [Pearl, Sprouse 2013; Tomida, Utsumi 2013; Wilcox et al. 2018; Wilcox et al. 2022]). Отметим один важный недостаток всех перечисленных экспериментов: во всех них моделировалось усвоение только лишь отношений между филлером и пробелом. При этом успешность усвоения способности порождать множество связанных с островами корректных грамматических структур не контролировалась. Так, модели не порождают неграмматичных структур при вопросительном выдвижении и других видах подъема, но не теряют ли они способность порождать все допустимые грамматичные структуры, связанные с такими передвижениями?

В последних из двух приведенных работ объектом исследования выступали нейронные сети (RNN, трансформерные модели) — модели, внутренние состояния которых непрозрачны. Исследование языковых способностей таких моделей напоминает изучение лингвистических компетенций человека, которые также во

[2] В биологии подобный механизм называется «эффектом Болдуина»: отбор определенных свойств (в нашем случае — грамматических запретов при передаче грамматики) происходит не напрямую как позволяющих либо не позволяющих воспроизводиться, а благодаря более простому усвоению этих свойств [Baldwin 1896].

многом представляют собой «черный ящик». Этот факт позволяет распространить на БЯМ экспериментальные методы, использующиеся в лингвистике для изучения компетенций естественных носителей.

Эксперименты по сравнению языковой способности у людей и машин могут принимать разную форму. Так, в исследовании [Lake, Baroni 2023] исследуется сравнительная способность людей и машин производить систематическую работу по построению композициональных структур[3]. В работе [Evanson et al. 2023] проводится эксперимент по моделированию усвоения грамматических параметров детьми и трансформерными моделями. Как отмечают авторы, при всем различии в «архитектуре» языковой способности усвоение отдельных признаков происходит схожим образом: существует иерархия языковых структур, по которой при обучении должны «восходить» как машины, так и люди.

На данный момент наиболее развитыми с точки зрения владения русским языком можно считать следующие три диалоговых трансформерных модели: ChatGPT[4], GigaChat[5], YandexGPT2[6]. Они демонстрируют правдоподобное поведение при ответе на запросы пользователей, в целом их ответы аналогичны человеческим с точки зрения содержания. Вопрос о том, насколько границы грамматичности всех трех моделей совпадают с таковыми у носителей-людей, представляет как теоретический, так и практический интерес. В данной работе мы сосредоточились на запрете на вынос из острова сочиненной структуры. Островные ограничения хорошо подходят для задачи исследования языковых компетенций моделей, работающих в диалоговом режиме. Во-первых, островные ограничения являются строгими — эти ограничения последовательно разделяются большинством носителей, и их нарушение ведет к неграмматичности. Взяв в качестве исходных предложений полностью грамматичные, мы сможем породить однозначно неприемлемые стимулы для нашего исследования. Во-вторых, нарушение островных ограничений происходит в том числе в ходе вопросительного передвижения составляющих за пределы острова. Одна из основных задач рассматриваемых нами языковых моделей — давать ответы на вопросы пользователей. Таким образом, вопросительные стимульные предложения представляют собой естественные входные данные для моделей.

Языковые модели проходят при создании тщательное тестирование как при помощи специфичных для каждого конкретного проекта методов, так и в рамках

[3] Отметим при этом, что с точки зрения исследования человека условия эксперимента, видимо, нельзя назвать достаточно корректными. Согласно стандартным представлениям порождающей грамматики, люди прибегают к врожденной способности к композициональности для усвоения грамматики того или иного языка в детстве, когда грамматика родного языка «выводится» из входного языкового материала автоматически и без сознательных усилий. В данном эксперименте, условия которого больше напоминают лингвистические олимпиадные задачи, люди поставлены перед необходимостью логически оценивать грамматичность тех или иных структур и осознанно строить деривации на искусственном «языке».

[4] 175 млрд параметров, трансформерная архитектура (см.: Introducing ChatGPT. *OpenAI*. https://openai.com/index/chatgpt/, дата обращения: 26.12.2023).

[5] 7 млрд параметров, трансформерная архитектура (см.: GigaChat расправляет плечи. Новая версия нейросетевой модели от Сбера. *Хабр*. https://habr.com/ru/companies/sberbank/articles/767492/, дата обращения: 26.12.2023).

[6] 7 млрд параметров, трансформерная архитектура (см.: YandexGPT 2 — большое обновление языковой модели Яндекса. *Хабр*. https://habr.com/ru/companies/yandex/articles/759306/, дата обращения: 26.12.2023).

общедоступных процедур оценок. Последние, такие как, например, SuperGLUE [Wang et al. 2019] или Russian SuperGLUE [Fenogenova et al. 2021], призваны оценивать способности моделей к логическому выводу, анализу информации, рассуждениям и т. д. По сути, большинство тестов и диагностик, если не все, связаны с проверкой тех или иных общекогнитивных навыков моделей. В то же время, проверкой собственно владения грамматикой, насколько нам известно, при создании моделей занимаются минимально. Почти наверняка можно утверждать, что чувствительность БЯМ к отрицательному языковому материалу прежде не подвергалась оценке и изучению — в этом состоит новизна настоящего исследования.

Материал и методы исследования

Мы протестировали чувствительность к острову сочинительной конструкции у трех языковых моделей на основе технологии GPT-3.5: ChatGPT, GigaChat и YandexGPT. К моделям применялись два теста; далее мы обсудим каждый из них.

В рамках первого теста проверялась способность моделей давать ответы на вопросы, в которых нарушены ограничения острова сочинительной конструкции. Пример такого вопроса приводится в (4): (4b) — вопрос к утвердительному предложению (4a). В (4b) вопросительная составляющая *кто* подвергается передвижению за пределы сочинительной конструкции, обозначенной скобками. Ожидаемый при игнорировании островных ограничений верный ответ на этот вопрос представлен в (4c) — далее мы будем называть такие ответы семантически верными. С точки зрения носителей естественного языка предложение (4b) неграмматично. Мы ожидаем, что люди будут испытывать существенные трудности при интерпретации этого вопроса и, следовательно, не смогут дать правильный ответ на него. Это ожидание следует из устройства модели языка, предложенной в минимализме (см. обсуждение выше). Неграмматичная структура не может быть создана синтаксическим модулем и, следовательно, в ситуации естественного порождения речи не может поступать на вход ни в фонетический (*PF*, *phonological form*), ни в семантический (*LF*, *logical form*) интерфейс. Наше предположение заключается в том, что если языковое поведение БЯМ приближено к поведению человека, то они также столкнутся с затруднениями при ответе на неграмматичный вопрос[7].

[7] Анонимный рецензент отметил спорность тезиса о том, что неграмматичность вопроса повлечет отказ носителя естественного языка от ответа на него. Близкая точка зрения встретилась нам и в литературе: в рецензии на монографию [Boeckx 2012] Д. Отт пишет о том, носители способны порождать интерпретацию для предложений с нарушением островных ограничений типа What does John like and oranges? ‘Что Джон любит и апельсины?’ [Ott 2014: 290]. Отт приходит к выводу, что возможность создания интерпретации в семантическом интерфейсе — свидетельство того, что такие предложения вовсе не являются неграмматичными, а их низкая приемлемость объясняется какими-то другими факторами. Отметим, однако, что этот тезис противоречит общепринятой точке зрения и требует пересмотра значительной части лингвистической теории. Обсуждению этого и некоторых других вопросов, связанных с обработкой неграмматичных предложений, посвящена статья [Leivada, Westergaard 2020].

Наше предположение о поведении носителей обосновывается не только предсказаниями лингвистической теории, но и результатами неформального опроса нескольких носителей русского языка. Безусловно, в дальнейшем необходимо проведение полноценного эксперимента с достаточным числом респондентов. На данном этапе в качестве дополнительного аргумента можно привести результаты исследования [Rankin et al. 2015]: носители английского языка выбирали ответ «не знаю» в 25 % случаев при ответе на неграмматичные вопросы, содержащие ошибку в согласовании / в по-

(4) a) [Дирижер Валерий Гергиев и симфонический оркестр Мариинского театра] станут специальными гостями фестиваля.
b) *Кто [дирижер Валерий Гергиев и ~~кто~~] станут специальными гостями фестиваля?
c) Симфонический оркестр Мариинского театра.

В первый тест вошли две затравки (англ. *prompt*), которые в общем виде представлены в примерах (5a) и (5b). Обе затравки содержат инструкцию «Дай максимально краткий ответ», чтобы сократить количество ответов, дословно повторяющих исходное утвердительное предложение, так как в таком случае невозможно определить, было ли проинтерпретировано извлечение из острова. Вторая затравка отличается от первой тем, что содержит более эксплицитную формулировку задачи.

(5) a) *{Утвердительное предложение. Вопрос, содержащий нарушение ограничения острова сочиненной конструкции}*. Дай максимально краткий ответ.
b) Исходное предложение такое: «*{Утвердительное предложение}*». Вопрос к нему: «*{Вопрос, содержащий нарушение ограничения острова сочиненной конструкции}*». Дай максимально краткий ответ.

Ответ на вопрос размечался как семантически верный, если представлял собой конъюнкт из исходного предложения, замененный вопросительным словом в вопросе, или же включал помимо этого конъюнкта любые фрагменты исходного предложения, кроме второго конъюнкта, ср. ответ YandexGPT в примере (6b) на запрос (6a).

(6) a) Команда передала им игрушки и необходимые медикаменты. Что команда передала им игрушки и? Дай максимально краткий ответ.
b) Команда передала им необходимые медикаменты.

Второй тест напрямую обращался к оценке приемлемости вопросов с нарушением островных ограничений. Как и в исследованиях по экспериментальному синтаксису, направленных на изучение языковой компетенции людей (напр.: [Лютикова, Герасимова 2021]), при формулировании инструкции мы не обращались к лингвистическому понятию грамматичности. Вместо этого модели должны были оценить корректность предложения. Шаблон затравки приводится в примере (7). Ответом на такой запрос, верно моделирующим языковое поведение человека, был бы отрицательный, так как носители языка однозначно определяют неграмматичность предложений, содержащих вопросительный вынос из сочинительной конструкции.

зиции лексического глагола (например, *Which animal catch the squirrels?*). Заметим, что в тех случаях, когда носители все же давали содержательный ответ, они на самом деле интерпретировали не подобные неграмматичные вопросы, а их «исправленные» варианты: при обработке носители меняли либо морфологическую форму глагола (*Which animal catches the squirrels?*), либо позицию лексического глагола (*Which animal do the squirrels catch?*). В случае острова сочиненной конструкции такие простые «исправления» невозможны: у неграмматичных предложений нет грамматичных аналогов, которые при этом были бы фонологически близки к исходным, как в случае обсуждавшегося выше примера. Для интерпретации неграмматичных предложений с нарушением островных ограничений необходима реконструкция синтаксической структуры до передвижения, которая при этом не может быть проведена по стандартным правилам.

(7) Корректно ли с точки зрения русского языка такое предложение: «*{Вопрос, содержащий нарушение ограничения острова сочиненной конструкции}*»?

Материалом нашего исследования стали 40 предложений, полученных из корпуса [Гращенков 2024]. Поскольку исходные предложения встретились в реальных текстах, мы можем связать значительные трудности с интерпретацией вопросов и низкие оценки приемлемости не с изначальными свойствами утвердительных предложений, а с нарушением островных ограничений[8]. При отборе предложений учитывались следующие факторы:

- тип составляющей, на уровне которой происходит сочинение: именная группа (NP) / группа прилагательного (AP) / предложная группа (PP) / глагольная группа (VP)[9];
- падеж извлекаемой составляющей (только для NP и AP): им. п., вин. п., род. п.;
- порядок слов в предложении (только для NP и AP в им. п.): подлежащее предшествует сказуемому (SV) / подлежащее следует за сказуемым (VS).

Каждое условие было представлено четырьмя лексикализациями; при выборе количества лексикализаций мы ориентировались на количество, принятое в экспериментальном синтаксисе. Вопросы формировались путем замены одного из конъюнктов или составляющей, входящей в состав конъюнкта, вопросительным словом и его передвижением на левую периферию предложения. В случае AP и PP на левую периферию также передвигались составляющие, затронутые эффектом крысолова. Каждому исходному предложению соответствовало два вопросительных: в одном из них вопросительным словом заменялся первый конъюнкт (или составляющая в составе конъюнкта), в другом — второй. Таким образом, всего в рамках каждой из инструкций было сделано по 80 запросов. При анализе результатов мы также учитывали фактор наличия «висящего» союза *и*. Поскольку *и* — сочинительный союз, для него нехарактерно выступать в абсолютном конце предложения, где нет лексического материала, который мог бы быть частью второго конъюнкта. Мы предполагаем, что крайне низкая частотность предложений с висящим и в реальных текстах должна повлиять на ответы моделей. Примеры предложений, вошедших в состав запросов по каждому условию, приводятся в табл. 1.

Для каждой модели мы подсчитывали долю семантически верных ответов в каждой из двух затравок первого теста, долю верных ответов, совпавших в обеих затравках, среднее количество верных ответов, долю ответов «корректно» во втором тесте, а также оценивали влияние обсуждавшихся выше факторов на результаты. Кроме того, мы оценивали работу моделей по параметрам логичности, устойчивости и зависимости от лексического материала. Для оценки логичности сравнивались результаты двух тестов. К логичным относились следующие комбинации ответов: 1) семантически верный, корректно; 2) семантически неверный, некорректно. В качестве нелогичных отмечались комбинации: 1) семантически верный, некорректно; 2) семантически неверный, корректно. Устойчивость оценивалась

[8] Мы также протестировали исходные предложения на корректность (см. затравку из второго теста), чтобы убедиться в том, что возможная некорректность вопросов связана именно с извлечением из острова.

[9] Все примеры с глагольными группами содержали сочинение VP с вершиной-инфинитивом: *Он пообещал [найти преступников] и [наказать нарушителей закона].*

Таблица 1. **Примеры предложений, вошедших в запросы (по одному примеру на каждое условие)**

№	Исходное предложение	Вопрос	Сочиненные оставляющие	Номер конъюнкта	Падеж вопросительной составляющей	Порядок слов	Висящий союз
1	Дирижер Валерий Гергиев и симфонический оркестр Мариинского театра станут специальными гостями фестиваля.	Кто и симфонический оркестр Мариинского театра станут специальными гостями фестиваля?	NP	Первый	Номинатив	SV	Нет
2	Дирижер Валерий Гергиев и симфонический оркестр Мариинского театра станут специальными гостями фестиваля.	Кто дирижер Валерий Гергиев и станут специальными гостями фестиваля?	NP	Второй	Номинатив	SV	Нет
3	В состав бургеров входили черная булочка и плавленый черный сыр.	Что в состав бургеров входили и плавленый черный сыр?	NP	Первый	Номинатив	VS	Нет
4	В состав бургеров входили черная булочка и плавленый черный сыр.	Что в состав бургеров входили черная булочка и?	NP	Второй	Номинатив	VS	Да
5	Команда передала им игрушки и необходимые медикаменты.	Что команда передала им и необходимые медикаменты?	NP	Первый	Аккузатив	NA	Нет
6	Команда передала им игрушки и необходимые медикаменты.	Что команда передала им игрушки и?	NP	Второй	Аккузатив	NA	Да
7	Они способствуют сплочению администрации и сотрудников.	Сплочению чего они способствуют и сотрудников?	NP	Первый	Генитив	NA	Нет
8	Они способствуют сплочению администрации и сотрудников.	Сплочению чего они способствуют администрации и?	NP	Второй	Генитив	NA	Да
9	Доброе и радостное известие пришло вчера.	Какое известие и радостное пришло вчера?	AP	Первый	Номинатив	SV	Нет
10	Доброе и радостное известие пришло вчера.	Какое известие доброе и пришло вчера?	AP	Второй	Номинатив	SV	Нет
11	У них отсутствовала бухгалтерская и финансовая отчетность.	Какая отчетность у них отсутствовала и финансовая?	AP	Первый	Номинатив	VS	Нет

12	У них отсутствовала бухгалтерская и финансовая отчетность.	Какая отчетность у них отсутствовала бухгалтерская и?	AP	Второй	Номинатив	VS	Да
13	Музыканты исполняют альтернативный и экспериментальный рок.	Какой рок музыканты исполняют и экспериментальный?	AP	Первый	Аккузатив	NA	Нет
14	Музыканты исполняют альтернативный и экспериментальный рок.	Какой рок музыканты исполняют альтернативный и?	AP	Второй	Аккузатив	NA	Да
15	На международном рынке идет восстановление оптовых и розничных цен.	Каких цен на международном рынке идет восстановление и розничных?	AP	Первый	Генитив	NA	Нет
16	На международном рынке идет восстановление оптовых и розничных цен.	Каких цен на международном рынке идет восстановление оптовых и?	AP	Второй	Генитив	NA	Да
17	Мой гардероб состоял из рваных вельветовых джинсов и из фланелевых рубашек.	Из чего мой гардероб состоял и из фланелевых рубашек?	PP	Первый	NA	NA	Нет
18	Мой гардероб состоял из рваных вельветовых джинсов и из фланелевых рубашек.	Из чего мой гардероб состоял из рваных вельветовых джинсов и?	PP	Второй	NA	NA	Да
19	Он пообещал найти преступников и наказать нарушителей закона.	Кого он пообещал найти и наказать нарушителей закона?	VP	Первый	NA	NA	Нет
20	Он пообещал найти преступников и наказать нарушителей закона.	Кого он пообещал найти преступников и наказать?	VP	Второй	NA	NA	Нет

NP (noun phrase) — именная группа, AP (adjective phrase) — группа прилагательного, PP (prepositional phrase) — предложная группа, VP (verb phrase) — глагольная группа, SV (subject + verb) — прямой порядок следования подлежащего и сказуемого, VS (verb + subject) — обратный порядок следования подлежащего и сказуемого, NA (not applicable) — не применимо.

при помощи сравнения результатов двух затравок в рамках первого теста. Ответ считался устойчивым, если: 1) был семантически верным в обоих запросах, 2) был семантически неверным в обоих запросах. Случаи, где верность ответа зависела от запроса, размечались как неустойчивые. Параметр зависимости от лексического материала показывает долю устойчиво верных ответов среди всех ответов, которые были верными при использовании хотя бы одной из затравок. Результаты работы всех трех моделей сравнивались между собой.

Результаты

ChatGPT

Общие характеристики:

- доля верных ответов. В рамках первого теста при обоих запросах количество семантически верных ответов — меньше половины. При использовании первой затравки доля верных ответов составляет 43,75 %, при использовании второй — 31,25 %. В среднем верные ответы давались в 37,5 % случаев. Доля верных ответов, совпавших при использовании двух затравок, составила 25 %;
- оценка корректности. Корректными были признаны 56,25 % вопросов с нарушением островных ограничений;
- логичность. Ответы логичны только примерно в половине случаев. При использовании результатов первой затравки из первого теста показатель логичности принимает значение 47,5 %, при использовании второй затравки — 52,5 %. Таким образом, более эксплицитная инструкция вызывает более логичное поведение модели, хотя различия незначительны;
- устойчивость. ChatGPT демонстрирует довольно высокую степень устойчивости: ответ на два запроса в рамках первого теста совпал в 75 % случаев;
- зависимость от лексического материала. Этот показатель также принимает относительно большое значение (50 %). Среди всех верных ответов половина — устойчиво верные.

Грамматические параметры:

- тип составляющей, на уровне которой происходит сочинение. Этот фактор оказывает влияние на долю верных ответов. В случае сочинения PP и VP доля семантически верных ответов превышает 50 % в обоих тестах, тогда как для NP и AP показатели существенно ниже. Этому факту находится объяснение в лингвистической теории. В литературе отмечается, что ограничение острова сочинительной конструкции более строгое, если при синтаксических процессах задействуется один из конъюнктов, чем в том случае, когда задействуется составляющая, входящая в состав одного из двух конъюнктов [Grosu 1973; Krejci 2020: 26]. В нашем материале в примерах с NP и AP вопросительными составляющими являются сами конъюнкты, тогда как в случае с PP и VP сочинение происходит на структурном уровне, превышающем вопросительную именную группу;
- падеж извлекаемой составляющей. В целом наблюдается больше семантически верных ответов и ответов «корректно», когда извлекаемые составля-

ющие выступают в им. п. Однако этот результат нельзя назвать последовательным: показатели зависят от категории составляющей;
- порядок слов в предложении. Результаты для фактора порядка слов непоследовательны. При использовании первой затравки доля верных ответов больше при порядке VS (68,75 % против 56,25 % для SV), тогда как при использовании второй затравки наблюдается обратная ситуация (18,75 % при VS и 31,25 % при SV). Существенно большее количество ответов «корректно» было получено при порядке SV (81,25 % против 31,25 %);
- вынос из первого/второго конъюнкта. При ответе на вопрос более приоритетным оказывается вынос из второго конъюнкта — для первой затравки 50 % верных ответов против 37,5 %, для второй затравки 35 % против 27,5 %. При оценке корректности вопроса, напротив, больше ответов «корректно» было дано для выноса из первого конъюнкта (67,5 % против 45 %);
- висящий союз. Не было найдено существенного влияния наличия висящего союза на долю верных ответов. Однако висящий союз ожидаемо негативно влияет на оценку корректности: для предложений без висящего союза было получено 67,92 % ответов «корректно», тогда как для предложений с ним — 33,33 %.

YandexGPT

Общие характеристики:
- доля верных ответов. Количество верных ответов, совпавших для первой и второй инструкций, составило 23,75 %, среднее количество верных ответов — 48,75 %. Наблюдается значительная разница в количестве правильных ответов на вопросы для первой (26,3 %) и второй (71,3 %) затравок;
- оценка корректности. Положительный ответ на вопрос о грамматичности, содержащийся в третьей инструкции, был получен для 53,8 % предложений;
- логичность. Для модели YandexGPT логичные пары ответов, где был дан правильный ответ и выставлена положительная оценка грамматичности, составили 60 % для первой затравки и 55 % для второй затравки;
- устойчивость. Устойчивые пары, для которых ответы на вопросы совпадают при использовании первой и второй инструкций, составили ровно 50 %;
- зависимость от лексического материала. Повторяемость лексического материала, т. е. отношение ответов на вопросы, верных для обеих затравок, ко всем верным ответам, принимает значение 32,2 %.

Грамматические параметры:
- тип составляющей, на уровне которой происходит сочинение. Универсальным для всех инструкций является тот факт, что модель наименее чувствительна к выносу из глагольной группы (VP). Для этой составляющей наблюдается наибольшее количество правильных ответов и положительных оценок грамматичности. При использовании второй и третьей затравок YandexGPT оказывается наиболее чувствительна к выносу группы прилагательного (AP), что выражается в наибольшем количестве неправильных ответов и отрицательных оценок грамматичности. При ответе на вопрос (первая и вторая затравки) наблюдается много неправильных ответов, если выносится именная группа (NP);

- падеж извлекаемой составляющей. Поведение модели оказывается непоследовательными для разных инструкций и разных типов составляющих;
- порядок слов в предложении. Данный фактор оказывает влияние на чувствительность к выносу из острова. Как для первой, так и для второй инструкции больше правильных ответов было получено при порядке VS, чем при порядке SV. При оценке приемлемости, напротив, больше положительных ответов было получено при порядке SV, чем при порядке VS;
- вынос из первого/второго конъюнкта. Результаты для первой и второй затравок совпадают. Больше правильных ответов наблюдается при выносе из первого конъюнкта, однако разница невелика: 32,5 % и 20 % для первого промпта, 75 % и 65,5 % для второго промпта. Для категориальных оценок грамматичности соотношение иное. Только 42,5 % предложений признаются грамматичным при выносе из первого конъюнкта, доля увеличивается до 65 % при выносе из второго конъюнкта;
- висящий союз. Было получено больше правильных ответов при отсутствии висящего союза как для первой, так для второй затравки. Положительная оценка приемлемости, наоборот, чаще выставлялась при наличии висящего союза.

GigaChat

Общие характеристики:

- доля верных ответов. Поведение GigaChat отличалось от предыдущих моделей небольшим количеством верных ответов на неграмматичные вопросы: 2,5 % для пересечения в двух экспериментах и 6,9 % в среднем для двух экспериментов. При использовании первой инструкции доля верных ответов (3,75 %) оказалась меньше, чем при использовании второй инструкции (10 %);
- оценка корректности. Прямой запрос о корректности показал, что корректным было признано 43,75 % вопросов;
- логичность. Логичность признания пар как не/корректных в вопросно-ответном тесте и при запросе корректности составила 55 %;
- устойчивость. Модель GigaChat продемонстрировала максимальный процент устойчивости при разных вариантах запросов: 91,3 % ответов были либо последовательно верными, либо последовательно неверными;
- зависимость от лексического материала. Процент последовательно верных ответов для обоих экспериментов относительно всех верных ответов — 22,2 %, что демонстрирует зависимость поведения модели от лексического материала, хотя и не самую высокую.

Грамматические параметры:

- тип составляющей, на уровне которой происходит сочинение. При ответе на вопрос наибольшее количество правильных ответов было получено при извлечении предложной группы (PP). Модель оказывается наименее чувствительна к извлечению из глагольной группы (VP) при ответе на вопрос с использованием второй затравки и при оценке корректности;

- падеж извлекаемой составляющей. При ответе на вопрос с помощью первой затравки и при оценке корректности наблюдается наименьшая чувствительность к выносу составляющей (NP, AP) в им. п.;
- порядок слов в предложении. Наибольшее количество правильных ответов было отмечено при выносе из структур с порядком слов VS (11,11 % для первого и 25,93 % для второго подхода) в противоположность выдвижению из структур SV (0 % для первого и второго подходов). Большее количество положительных оценок корректности напротив было получено при порядке SV (87,5 %), а не при VS (25 %);
- вынос из первого/второго конъюнкта. Можно диагностировать существенно более частый вынос из второго конъюнкта (81,8 %) в случае ответа на вопрос. Однако предложения чаще оценивались как корректные при извлечении из второго, чем из первого конъюнкта (55 % и 32 % соответственно);
- висящий союз. Наличие висящего союза способствует уменьшению чувствительности к выносу из острова при ответе на вопрос (11,11 % и 0 % для первой затравки, 25,93 % и 1,89 % для второй затравки). Неграмматичность предложений чаще игнорировалась при отсутствии висящего союза (7,41 % и 62,26 %).

Обсуждение

Подведем итоги проведенного тестирования и обобщим результаты, полученные для трех моделей: ChatGPT, YandexGPT и GigaChat. Наибольшее количество семантически верных ответов на вопрос с нарушением ограничения острова сочинительной конструкции дала модель YandexGPT (26,25 % и 71,25 % для первой и второй затравок), наименьшее — модель GigaChat (3,75 % и 10 % для первой и второй затравок). Можно заметить, что формулировка инструкции оказывает наибольшее влияние на способность модели YandexGPT отвечать на вопросы, содержащие вынос из острова сочинительной конструкции. При этом больше всего неграмматичных, с точки зрения людей, предложений были оценены как корректные моделью ChatGPT (56,25 %), меньше всего — моделью GigaChat (43,75 %). Можно сделать вывод, что GigaChat проявляет наибольшую чувствительность к островным ограничениям. Это свойство приближает его лингвистическую компетенцию к языковым способностям человека. Мы можем быть уверенными в том, что оценка предложений как некорректных определяется нарушением островного ограничения сочиненной конструкции, а не лексическими факторами: модель признала все исходные предложения корректными.

Мы также оценили естественность языкового поведения моделей по параметрам логичности и устойчивости. Мы исходим из предположения о том, что носители естественного языка, с одной стороны, не смогут дать верный ответ на вопрос, который считают неграмматичным, с другой стороны, будут давать совпадающие ответы на один и тот же вопрос[10]. Показатель логичности был призван измерить последовательность модели при ответе на вопрос и оценке корректности предло-

[10] Впрочем, людям тоже свойственна некоторая доля непоследовательности — см. работу [Герасимова и др. 2024] о последовательности при оценке стимульных предложений в ходе синтаксических экспериментов.

жений. Для моделей YandexGPT и GigaChat наиболее высокое значение показателя наблюдается при использовании первой затравки, содержащей минимальную инструкцию. Для ChatGPT больше «логичных» пар оказывается при использовании второй затравки, более эксплицитно формулирующей задание. В целом поведение моделей логично только примерно в половине случаев: больший из двух результат — 52,5 % для ChatGPT, 55 % для GigaChat, 60 % для YandexGPT. Мы можем сделать вывод, что поведение YandexGPT на наших данных наиболее логично. Как можно видеть, модели ChatGPT и GigaChat часто оказывались не готовы отвечать на вопросы, которые сами считали корректными. Показатель устойчивости демонстрирует, насколько совпадают ответы модели при переходе между первой и второй затравками. Наибольшее значение устойчивости демонстрирует GigaChat (91,25 %), наименьшее — YandexGPT (50 %). Можно заметить, что поведение моделей существенно различается по этому параметру. Показатель зависимости от лексического материала также демонстрирует существенную вариативность в поведении моделей. Этот показатель определяется тем, сколько ответов из всех семантически верных были устойчивыми. Он оказался наибольшим для ChatGPT (50 %) и наименьшим для GigaChat (22,22 %).

Модели демонстрируют одинаковое поведение только в отношении части грамматических свойств. К ним относятся тип извлекаемой составляющей и порядок слов. Так, для всех моделей легче достигается извлечение составляющей, входящей в состав конъюнкта, а не целого конъюнкта, — эта черта находит параллель в лингвистической теории. Модель ChatGPT оказывается стабильно менее чувствительна к выносу из PP и VP, что выражается в большем количестве правильных ответов и положительных оценок. Для YandexGPT и GigaChat аналогичная закономерность наблюдается только при выносе из глагольной группы. Другая закономерность состоит в том, что большее количество правильных ответов при извлечении из NP и AP в субъектной позиции наблюдается в случае обратного порядка подлежащего и сказуемого (VS). Напротив, больше положительных оценок приемлемости было получено при прямом порядке подлежащего и сказуемого (SV). Можно отметить, что влияние порядка слов на чувствительность к выносу из острова оказывается одинаковым для всех моделей, однако зависит от задачи (ответ на вопрос vs оценка корректности).

Роль других грамматических параметров совпадает для одних моделей и различается для других. Так, различными оказываются предпочтения моделей к тому, из какого конъюнкта происходит извлечение — из первого или из второго. Висящий союз также может как способствовать верным ответам, так и, напротив, затруднять их. Для моделей ChatGPT и GigaChat наблюдается большее количество верных ответов при выносе из второго конъюнкта и наличии висящего союза, для YandexGPT — при выносе из первого конъюнкта и отсутствии висящего союза. При этом интересно, что больше положительных оценок корректности наблюдается в противоположном случае: для ChatGPT и GigaChat — при выносе из первого конъюнкта и отсутствии висящего союза, для YandexGPT — при выносе из второго конъюнкта и наличии висящего союза. Таким образом, ChatGPT и GigaChat демонстрируют схожее поведение в противоположность модели YandexGPT. Отметим также, что противоположные грамматические предпочтения в двух задачах (ответ на вопрос vs оценка корректности) скорее являются неестественными с точки зре-

ния носителей языка и представляют собой пример различий в языковом поведении БЯМ и человека.

Падеж извлекаемой составляющей в случае NP и AP оказывает самое непоследовательное влияние на чувствительность моделей к острову сочинительной конструкции. ChatGPT демонстрирует наибольшее количество правильных ответов и положительных оценок корректности при выносе составляющей в именительном падеже, хотя для NP и AP ответы различаются. Для GigaChat эта закономерность также наблюдается при оценке корректности, но при ответе на вопрос сохраняется только для первой затравки. YandexGPT оказывается наименее чувствительна к извлечению составляющей в именительном падеже только при ответе на вопрос с использованием первой инструкции. Следовательно, наблюдаемая закономерность не является универсальной, мы видим большую вариативность в поведении моделей.

Общим выводом может быть то, что структуры с нарушением ограничения сочинительного острова не являются абсолютно неграмматичными ни для одной из моделей. Степень неграмматичности может варьировать в зависимости от выбора модели и диагностики, но нигде не приближается к человеческой: для носителей естественного языка мы ожидали бы полного отсутствия семантически верных ответов и оценок «корректно». Возможно, в случае вопросно-ответного тестирования результаты частично объясняются постановкой задачи при разработке моделей. Основная функция рассмотренных нами моделей — максимально эффективно давать ответы на запросы пользователей, пренебрегая опечатками и грамматической небезупречностью введенных данных. Согласно полученным при обучении инструкциям, не ответить на запрос пользователя — неправильный для модели результат. Отметим, что такого конфликта не должно возникать при запросе, в котором от модели требуется оценить корректность предложения. Тем не менее мы видим, что оценку «некорректно» неграмматичные с точки зрения носителей предложения получают только примерно в половине случаев.

Заключение

В данной работе мы протестировали лингвистические способности трех русскоязычных БЯМ, работающих в диалоговом режиме: ChatGPT, GigaChat и YandexGPT. Объектом исследования послужила чувствительность моделей к острову сочинительной конструкции. В русском языке этот остров является сильным — вынос составляющих за его пределы однозначно оценивается носителями языка как неграмматичный. На наш взгляд, этот феномен представляет собой хороший материал для поиска различий в лингвистической компетенции человека и языковых моделей. С одной стороны, наличие грамматичного исходного предложения позволяет убедиться в том, что предложение с нарушением островных ограничений неприемлемо по грамматическим причинам, а не по лексическим. С другой стороны, деривирование нарушения островных ограничений с помощью вопросительного передвижения позволяет получить стимулы, которые являются естественными входными данными для диалоговых моделей.

Для проверки чувствительности моделей к сочинительному острову использовались два теста. В ходе первого теста проверялась способность моделей отвечать

на вопросы, в которых есть нарушение островного ограничения. Естественным исходом с точки зрения носителя языка была бы неспособность проинтерпретировать предложение и, следовательно, дать семантически верный ответ на такой вопрос. Поведение моделей отличается от поведения человека: количество семантически верных ответов не только не равно нулю, но и очень велико в некоторых случаях (наибольший показатель — 71,25 % у YandexGPT). Возможным объяснением такого результата может служить заложенная при разработке моделей нацеленность на порождение ответа даже в условиях неидеальных входных данных. В первом тесте использовались две затравки, различающихся эксплицитностью инструкции. Реакция моделей на формулировку инструкции различна: в случае ChatGPT большее число семантически верных ответов было получено при использовании более подробной инструкции, тогда как GigaChat и YandexGPT демонстрируют обратный результат.

Целью второго теста была прямая проверка того, считают ли модели предложения с нарушением островных ограничений возможными с точки зрения русского языка. Хотя нас интересовала грамматичность стимульных предложений, в инструкции мы использовали понятие корректности, чтобы сделать запрос как можно более «понятным» для моделей. Этот тест также показал, что лингвистическая компетенция БЯМ отличается от человеческой. Каждая из моделей оценила как некорректные только около половины предложений, деривированных с нарушением островных ограничений.

В рамках исследования мы разработали несколько метрик для характеристики поведения моделей. Логичность измеряла единообразие ответов моделей при переходе от одного теста к другому, устойчивость — последовательность ответов при применении различных затравок в рамках одного теста, зависимость от лексического материала — долю устойчивых ответов среди всех семантически верных. По этим показателям рассмотренные нами модели также демонстрируют результаты, неожиданные с точки зрения носителей естественного языка.

Мы проанализировали влияние различных грамматических факторов на долю верных ответов и оценку корректности предложений. Рассмотренные нами модели не только демонстрируют поведение, отличное от человеческого, но и обладают различающимися грамматическими предпочтениями: только часть факторов единообразно влияет на ответы моделей. При этом ChatGPT и GigaChat скорее ведут себя схожим образом, в противоположность YandexGPT.

Хотя вопрос правдоподобия поведения языковых моделей изучался и другими исследователями, наша работа не только вносит вклад в развитие методологии исследований по данной проблематике, но и расширяет базу эмпирических обобщений. Мы рассмотрели небольшой фрагмент грамматической системы на представительной выборке стимульных предложений, полученных с помощью обработки корпусного материала. Поскольку задача диалоговых моделей — воспроизводить поведение человека, наши инструкции были сформулированы так, как будто мы обращались к носителю языка. Данный подход позволяет приблизить исследования языковых моделей к исследованиям по экспериментальному синтаксису. В качестве одного из направлений продолжения работы мы видим изучение поведения людей и диалоговых языковых моделей на одном и том же материале. Подобная постановка задачи сделает возможным прямое сравнение лингвистической ком-

петенции человека с языковой способностью нейронных сетей. Распространение предложенной нами методологии на другие лингвистические феномены позволит описать грамматику БЯМ.

Результаты нашего исследования свидетельствуют: такая грамматика значительно отличается от человеческой. Таким образом, утверждения о том, что поведение БЯМ неотличимо от поведения людей, кажутся нам преждевременными.

Литература

Герасимова и др. 2024 — Герасимова А. А., Лютикова Е. А., Паско Л. И. Языковая компетенция сквозь призму грамматической вариативности. Часть 1. Теоретические и методологические соображения. *Вестник Московского университета. Серия 9. Филология.* 2024, (4): 9–22.

Гращенков 2024 — Гращенков П. В. RuConst: Синтаксический корпус русского языка с разметкой по непосредственным составляющим. *Вестник Московского университета. Серия 9. Филология.* 2024, (3): 94–112.

Зализняк, Падучева 1979 — Зализняк А. А., Падучева Е. В. Синтаксические свойства местоимения *который*. В кн.: *Категория определенности-неопределенности в славянских и балканских языках*: сб. ст. Николаева Т. М. (отв. ред.). М.: Наука, 1979. С. 289–329.

Лютикова, Герасимова 2021 — Лютикова Е. А., Герасимова А. А. (ред.). *Русские острова в свете экспериментальных данных.* М.: Буки Веди, 2021.

Моргунова 2021 — Моргунова Е. В. Островные конструкции в русском языке. В кн.: Русские острова в свете экспериментальных данных. Лютикова Е. А., Герасимова А. А. (ред.). М.: Буки Веди, 2021. С. 35–55.

Baldwin 1896 — Baldwin M. J. A New Factor in Evolution. *The American Naturalist.* 1896, 30 (354): 441–451.

Boeckx 2012 — Boeckx C. Syntactic islands. Cambridge: Cambridge University Press, 2012.

Chomsky 2004 — Chomsky N. Beyond explanatory adequacy. In: Structures and Beyond: The Cartography of Syntactic Structures. Belletti A. (ed.). Oxford: Oxford University Press, 2004. P. 104–131.

Chomsky 2013 — Chomsky N. Problems of Projection. *Lingua.* 2013, (130): 33–49.

Cinque 1990 — Cinque G. *Types of A' Dependencies.* Cambridge: MIT Press, 1990.

Evanson et. al. 2023 — Evanson L., Lakretz Y., King J.-R. Language acquisition: do children and language models follow similar learning stages? *Findings of the Association for Computational Linguistics: ACL 2023.* 2023: 12205–12218.

Fenogenova et al. 2021 — Fenogenova A., Shavrina T., Kukushkin A., Tikhonova M., Emelyanov A., Malykh V., Mikhailov V., Shevelev D., Artemova E. Russian SuperGLUE 1.1: Revising the Lessons not Learned by Russian NLP-models. In: *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2021"*. 2021. P. 267–277.

Grosu 1973 — Grosu A. On the nonunitary nature of the Coordiate Structure Constraint. *Linguistic Inquiry*. 1973, 4 (1): 88–92.

Krejci 2020 — Krejci B. *Syntactic and semantic perspectives on first conjunct agreement in Russian*. PhD thesis. Stanford: Stanford University, 2020.

Lake, Baroni 2023 — Lake B. M., Baroni M. Human-like systematic generalization through a meta-learning neural network. *Nature*. 2023, (623): 115–121.

Lakoff 1986 — Lakoff G. Frame semantic control of the Coordinate Structure Constraint. *Proceedings of the Chicago Linguistic Society.* 1986, (22): 152–167.

Leivada, Westergaard 2020 — Leivada E., Westergaard M. Acceptable ungrammatical sentences, unacceptable grammatical sentences, and the role of the cognitive parser. *Frontiers in Psychology.* 2020, (11): 364.

Ott 2014 — Ott D. Syntactic islands by Cedric Boeckx (review). *Language.* 2014, (90): 287–291.

Pearl, Sprouse 2013 — Pearl L., Sprouse J. Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. *Language Acquisition.* 2013, 20 (1): 23–68.

Phillips 2013a — Phillips C. On the nature of island constraints. I: Language processing and reductionist accounts. In Sprouse J., Hornstein N. (eds). *Experimental syntax and island effects.* Cambridge: Cambridge University Press, 2013. P. 64–108.

Phillips 2013b — Phillips C. On the nature of island constraints. II: Language processing and reductionist accounts. In Sprouse J., Hornstein N. (eds). *Experimental syntax and island effects.* Cambridge: Cambridge University Press, 2013. P. 132–157.

Rankin et al. 2015 — Rankin T., Grosso S., Reiterer S. Effects of L1 co-activation on the processing of L2 morpho-syntax in German-speaking learners of English. In: Stringer D. et al. (eds). *Proceedings of the 13th Generative Approaches to Second Language Acquisition Conference (GASLA 2015)*. 2015. P. 196–207.

Ross 1967 — Ross J. R. *Constraints on variables in syntax.* PhD thesis. Cambridge, Massachusetts: Massachusetts Institute of Technology, 1967.

Tomida, Utsumi 2013 — Tomida Y., Utsumi A. A connectionist model for acquisition of syntactic islands. *Procedia — Social and Behavioral Sciences.* 2013, (97): 90–97.

Wang et al. 2019 — Wang A., Pruksachatkun Y., Nangia N., Singh A., Michael J., Hill F., Levy O., Bowman S. R. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems*. 2019. P. 3261–3275.

Wilcox et al. 2018 — Wilcox E. G., Levy R., Takashi M., Futrell R. What do RNN language models learn about filler–gap dependencies? In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP.* Brussels, 2018. P. 211–221.

Wilcox et al. 2022 — Wilcox E. G., Futrell R., Levy R. Using computational models to test syntactic learnability. *Linguistic Inquiry*. 2022, Special Collection: CogNet: 1–44.

Williams 1978 — Williams E. Across-the-board rule application. *Linguistic Inquiry.* 1978, 9 (1): 31–43.

Статья поступила в редакцию 26 ноября 2023 г.
Статья рекомендована к печати 14 мая 2024 г.

Pavel V. Grashchenkov

Lomonosov Moscow State University,
1, Leninskie Gory, Moscow, 119991, Russia
https://orcid.org/0000-0001-9754-2452
pavel.gra@gmail.com

Kseniia A. Studenikina

Lomonosov Moscow State University,
1, Leninskie Gory, Moscow, 119991, Russia
https://orcid.org/0000-0002-4098-7167
xeanst@gmail.com

Lada I. Pasko

Lomonosov Moscow State University,
1, Leninskie Gory, Moscow, 119991, Russia
https://orcid.org/0000-0002-0533-809X
paskolada@yandex.ru

Coordinate structure constraint in the linguistic competence of large language models*

For citation: Grashchenkov P. V., Studenikina K. A., Pasko L. I. Coordinate structure constraint in the linguistic competence of large language models. *Vestnik of Saint Petersburg University. Language and Literature.* 2024, 21 (3): 668–688. https://doi.org/10.21638/spbu09.2024.309 (In Russian)

A syntactic island is a construction extraction from which leads to ungrammaticality. Island constraints are generally demonstrated through the impossibility of the A′-movement, e. g.

* This work was done with the support of Lomonosov Moscow State University Program of Development, project no. 23-III02-10 “Linguistic competence of natural language speakers and neural network models”.

wh-movement. Considering extraction from an island as ungrammatical is common to all native speakers. In terms of natural language understanding and generation, the competence of large language models (LLM) is almost indistinguishable from the human one. However, the difference between the grammatical constraints of the native speakers and LLM are still studied insufficiently. If the LLM grammar is set up similar to the human one, they will demonstrate high sensitivity to island constraints. The current study aims to compare the language competence of the native speakers and LLM based on the coordinate structure islands. The three dialogue systems — ChatGPT, YandexGPT and GigaChat — were examined via two tests. The first one investigates whether the model is able to give a semantically correct answer to the question with violation of island constraints. The second test directly accesses the grammaticality judgements. The results clearly show that the LLM language competence differs from the human one. The observed models regularly answer the questions violating island constraints correctly and consider them grammatical. YandexGPT turns out to be more consistent, while ChatGPT and GigaChat frequently give incorrect answers to the questions which they judge acceptable. The influence of the stimuli's grammatical features depends on the model: the island sensitivity of ChatGPT and GigaChat is determined by the same features in contrast to YandexGPT. Thus, the results call into question the fact that LLM language competence is close to the human one.

Keywords: large language models, natural language processing, Russian, syntactic island, syntax.

References

Герасимова и др. 2024 — Gerasimova A. A., Lyutikova E. A., Pasko L. I. Linguistic Competence Through the Lens of Grammatical Variation. Part 1. Conceptual and Methodological Considerations. *Vestnik Moskovskogo universiteta*. *Ser. 9. Filologiia*. 2024, (4): 9–22. (In Russian)

Гращенков 2024 — Grashchenkov P. V. RuConst: A Treebank for Russian. *Vestnik Moskovskogo universiteta*. Ser. 9. Filologiia. 2024, (3): 94–112. (In Russian)

Зализняк, Падучева 1979 — Zalizniak A. A., Paducheva E. V. Syntactic properties of the pronoun *kotoryj*. In: *Kategoriya opredelennosti-neopredelennosti v slavyanskih i balkanskih yazykah*: sbornik statei. Nikolaeva T. M. (ed.). Moscow: Nauka Publ., 1979. P. 289–329. (In Russian)

Лютикова, Герасимова 2021 —*Russian islands in the light of experimental data*. Liutikova E. A., Gerasimova A. A. (eds). Moscow: Buki Vedi Publ., 2021. (In Russian)

Моргунова 2021 — Morgunova E. V. Island constraints in Russian. In: *Russian islands in the light of experimental data*. Liutikova E. A., Gerasimova A. A. (eds). Moscow: Buki Vedi Publ., 2021. P. 35–55. (In Russian)

Baldwin 1896 — Baldwin M. J. A New Factor in Evolution. *The American Naturalist.* 1896, 30 (354): 441–451.

Boeckx 2012 — Boeckx C. *Syntactic islands.* Cambridge: Cambridge University Press, 2012.

Chomsky 2004 — Chomsky N. Beyond explanatory adequacy. In: *Structures and Beyond: The Cartography of Syntactic Structures*. Belletti A. (ed.). Oxford: Oxford University Press, 2004. P. 104–131.

Chomsky 2013 — Chomsky N. Problems of Projection. *Lingua.* 2013, (130): 33–49.

Cinque 1990 — Cinque G. *Types of A' Dependencies.* Cambridge: MIT Press, 1990.

Evanson et. al. 2023 — Evanson L., Lakretz Y., King J.-R. Language acquisition: do children and language models follow similar learning stages? *Findings of the Association for Computational Linguistics: ACL 2023.* 2023: 12205–12218.

Fenogenova et al. 2023 — Fenogenova A., Shavrina T., Kukushkin A., Tikhonova M., Emelyanov A., Malykh V., Mikhailov V., Shevelev D., Artemova E. Russian SuperGLUE 1.1: Revising the Lessons not Learned by Russian NLP-models. In: *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2021"*. 2021. P. 267–277.

Grosu 1973 — Grosu A. On the nonunitary nature of the Coordiate Structure Constraint. *Linguistic Inquiry*. 1973, 4 (1): 88–92.

Krejci 2020 — Krejci B. *Syntactic and semantic perspectives on first conjunct agreement in Russian*. PhD thesis. Stanford: Stanford University, 2020.

Lake, Baroni 2023 — Lake B. M., Baroni M. Human-like systematic generalization through a meta-learning neural network. *Nature*. 2023, (623): 115–121.
Lakoff 1986 — Lakoff G. Frame semantic control of the Coordinate Structure Constraint. *Proceedings of the Chicago Linguistic Society.* 1986, (22): 152–167.
Leivada, Westergaard 2020 — Leivada E., Westergaard M. Acceptable ungrammatical sentences, unacceptable grammatical sentences, and the role of the cognitive parser. *Frontiers in Psychology.* 2020, (11): 364.
Ott 2014 — Ott D. Syntactic islands by Cedric Boeckx (review). *Language.* 2014, (90): 287–291.
Pearl, Sprouse 2013 — Pearl L., Sprouse J. Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. *Language Acquisition*. 2013, 20 (1): 23–68.
Phillips 2013a — Phillips C. On the nature of island constraints. I: Language processing and reductionist accounts. In: *Experimental syntax and island effects.* Sprouse J., Hornstein N. (eds). Cambridge: Cambridge University Press, 2013. P. 64–108.
Phillips 2013b — Phillips C. On the nature of island constraints. II: Language processing and reductionist accounts. In: *Experimental syntax and island effects.* Sprouse J., Hornstein N. (eds). Cambridge: Cambridge University Press, 2013. P. 132–157.
Rankin et al. 2015 — Rankin T., Grosso S., Reiterer S. Effects of L1 co-activation on the processing of L2 morpho-syntax in German-speaking learners of English. In: *Proceedings of the 13th Generative Approaches to Second Language Acquisition Conference (GASLA 2015)*. Stringer D. et al. (eds). 2015. P. 196–207.
Ross 1967 — Ross J. R. *Constraints on variables in syntax.* PhD thesis. Cambridge, Massachusetts: Massachusetts Institute of Technology, 1967.
Tomida, Utsumi 2013 — Tomida Y., Utsumi A. A connectionist model for acquisition of syntactic islands. *Procedia — Social and Behavioral Sciences*. 2013, (97): 90–97.
Wang et al. 2019 — Wang A., Pruksachatkun Y., Nangia N., Singh A., Michael J., Hill F., Levy O., Bowman S. R. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems*. 2019. P. 3261–3275.
Wilcox et al. 2018 — Wilcox E. G., Levy R., Takashi M., Futrell R. What do RNN language models learn about filler-gap dependencies? In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP.* Brussels, 2018. P. 211–221.
Wilcox et al. 2022 — Wilcox E. G., Futrell R., Levy R. Using computational models to test syntactic learnability. *Linguistic Inquiry*. 2022, Special Collection: CogNet: 1–44.
Williams 1978 — Williams E. Across-the-board rule application. *Linguistic Inquiry.* 1978, 9 (1): 31–43.

Received: November 26, 2023
Accepted: May 14, 2024