ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ
INTELLIGENT SYSTEMS

УДК 004.8
https://doi.org/10.37661/1816-0301-2022-19-1-96-110

Обзорная статья
Review Paper

Применение модели освоения языка к решению задачи обработки малых языков

Д. И. Качков

Белорусский государственный университет,
пр. Независимости, 4, Минск, 220030, Беларусь
E-mail: dmitriydikanskiy@gmail.com

Аннотация
Решается задача построения компьютерной модели малого языка. Ее актуальность обусловлена необходимостью устранения информационного неравенства между носителями различных языков, востребованностью новых инструментов для исследования малоизученных языков и инновационных подходов к моделированию языка в условиях дефицита ресурсов, необходимостью поддержки и развития языков малых народов.
При решении задачи обработки малых языков на этапе описания проблемной ситуации преследуются три основные цели: обоснование проблемы моделирования языка в условиях дефицита ресурсов как особой задачи в сфере моделирования естественных языков, обзор литературы по соответствующей теме и разработка концепции модели усвоения языка с относительно малым числом доступных ресурсов.
Используются методы компьютерного моделирования с применением нейронных сетей, обучение с частичным привлечением учителя и обучение с подкреплением.
В работе приведен обзор литературы, посвященной моделированию изучения лексики, морфологии и грамматики родного языка ребенком. На основании современных представлений о ходе изучения языка предложена архитектура системы обработки малого языка, которая при обучении опирается на компьютерное моделирование онтогенеза. Выделены основные компоненты системы и принципы их взаимодействия. В основе системы лежит модуль, построенный на базе современных диалоговых языковых моделей и обученный на каком-либо крупном языке, например английском. При обучении используется промежуточный слой, который представляет высказывания в некотором абстрактном виде, например, в символах формальной семантики. Соотношение между формальной записью высказываний и их переводом на целевой малый язык изучается методом моделирования процесса усвоения лексики и грамматики языка ребенком. Отдельный компонент имитирует неязыковой контекст, в котором происходит изучение языка.
В статье исследуется задача моделирования малых языков. Дано подробное обоснование актуальности моделирования малых языков: показана социальная значимость этой проблемы, польза ее решения для лингвистики, этнографии, этнологии и культурной антропологии. Отмечена неэффективность подходов, применяемых к крупным языкам, в условиях дефицита ресурсов. Предложена модель изучения языка с помощью имитации онтогенеза, которая опирается как на полученные результаты в области компьютерного моделирования, так и на данные психолингвистики.

Ключевые слова: информационные технологии, языковые модели, обработка малого языка, усвоение языка, обучение с подкреплением, нейронные сети, архитектура Transformer

Для цитирования. Качков, Д. И. Применение модели освоения языка к решению задачи обработки малых языков / Д. И. Качков // Информатика. – 2022. − Т. 19, № 1. – С. 96–110.
https://doi.org/10.37661/1816-0301-2022-19-1-96-110

Конфликт интересов. Автор заявляет об отсутствии конфликта интересов.

Поступила в редакцию | Received 18.11.2021
Подписана в печать | Accepted 08.12.2021
Опубликована | Published 29.03.2022

Applying the language acquisition model to the solution small language processing tasks

Dzmitry I. Kachkou

Belarusian State University,
av. Nezavisimosti, 4, Minsk, 220030, Belarus
E-mail: dmitriydikanskiy@gmail.com

Abstract
The problem of building a computer model of a small language was under solution. The relevance of this task is due to the following considerations: the need to eliminate the information inequality between speakers of different languages; the need for new tools for the study of poorly understood languages, as well as innovative approaches to language modeling in the low-resource context; the problem of supporting and developing small languages.
There are three main objectives in solving the problem of small natural language processing at the stage of describing the problem situation: to justify the problem of modeling language in the context of resource scarcity as a special task in the field of natural languages processing, to review the literature on the relevant topic, to develop the concept of language acquisition model with a relatively small number of available resources.
Computer modeling techniques using neural networks, semi-supervised learning and reinforcement learning were involved.
The paper provides a review of the literature on modeling the learning of vocabulary, morphology, and grammar of a child's native language. Based on the current understanding of the language acquisition and existing computer models of this process, the architecture of the system of small language processing, which is taught through modeling of ontogenesis, is proposed. The main components of the system and the principles of their interaction are highlighted. The system is based on a module built on the basis of modern dialogical language models and taught in some rich-resources language (e.g., English). During training, an intermediate layer is used which represents statements in some abstract form, for example, in the symbols of formal semantics. The relationship between the formal recording of utterances and their translation into the target low-resource language is learned by modeling the child's acquisition of vocabulary and grammar of the language. One of components stands for the non-linguistic context in which language learning takes place.
This article explores the problem of modeling small languages. A detailed substantiation of the relevance of modeling small languages is given: the social significance of the problem is noted, the benefits for linguistics, ethnography, ethnology and cultural anthropology are shown. The ineffectiveness of approaches applied to large languages in conditions of a lack of resources is noted. A model of language learning by means of ontogenesis simulation is proposed, which is based both on the results obtained in the field of computer modeling and on the data of psycholinguistics.

Keywords: information technology, language models, low-resource language processing, language acquisition, reinforcement learning, neural networks, Transformer architecture

For citation. Kachkou D. I. *Applying the language acquisition model to the solution small language processing tasks*. Informatika [*Informatics*], 2022, vol. 19, no. 1, pp. 96−110 (In Russ.).
https://doi.org/10.37661/1816-0301-2022-19-1-96-110

Conflict of interest. The author declare of no conflict of interest.

Введение. Решение проблемы автоматической обработки естественного языка является одним из наиболее актуальных направлений в области искусственного интеллекта. Современные языковые модели, как правило, основанные на архитектуре Transformer, позволяют достичь впечатляющих результатов в решении разнообразных задач, связанных с обработкой и пониманием текстов. Обучение таких моделей происходит в два этапа: первый требует огромного корпуса неразмеченных данных, второй – небольшой выборки размеченных данных, связанных с конкретной задачей.

Как показывают исследования, для раскрытия всего потенциала языковых моделей корпус неразмеченных текстов должен насчитывать миллиарды слов. Очевидно, что подобные обучающие выборки можно построить только для крупных языков: английского, русского, китайского, которые имеют развитую литературную традицию и широко используются в Интернете. Между тем автоматическая обработка малых языков – не менее важная задача для современной науки, решение которой позволит внести весомый вклад как в сохранение исчезающих языков, так и в компьютерную лингвистику.

Основные подходы, выработанные в рамках проблемы обработки малых языков, адаптируют крупные языковые модели к условиям дефицита доступных ресурсов. В первую очередь это методы искусственной генерации дополнительных текстов на целевом языке и трансфер знаний от модели крупного языка к модели малого языка.

Другая сфера компьютерной лингвистики – моделирование онтогенеза языка, т. е. процесса усвоения языка ребенком. Как правило, подобные модели используются для исследования гипотез, связанных с процессом освоения родного языка.

В настоящей работе выдвигается гипотеза о применимости метода моделирования онтогенеза к задаче моделирования малых языков.

1. Малые языки. Автоматическая обработка языков с дефицитом ресурсов (англ. low-resource languages) сформировалась как отдельная проблема в области компьютерной лингвистики. Исследователи активно ищут технологии и подходы, которые позволят повысить эффективность решения языковых задач в условиях малого объема доступной обучающей выборки.

Остановимся подробнее на термине «языки с дефицитом ресурсов». Можно выделить следующие категории ресурсов [1]:

1. Данные, размеченные для решаемой задачи. Размеченные данные необходимы для проведения обучения с учителем, поэтому отсутствие их в достаточном количестве становится значительной проблемой при решении поставленной задачи. Стоит отметить, что разметка данных нередко проводится вручную специалистами в данном языке и (или) в данной предметной области. Такие специалисты могут отсутствовать, в частности, по причине экзотичности рассматриваемого языка. Отсутствие соответствующего эксперта дополнительно осложняет решение задачи обработки языка.

2. Неразмеченные тексты на целевом языке или на целевую тему. Во многих современных моделях применяется «обучение с частичным привлечением учителя» (semi-supervised learning) [2]: на первом этапе происходит обучение без учителя на большом корпусе неразмеченных данных, на втором – дообучение под конкретную задачу на небольшом количестве размеченных примеров. Первый этап обучения требует больших корпусов текстов. Если корпус достаточного объема оказывается недоступен, построение эффективных языковых моделей на базе Transformer значительно затрудняется.

3. Вспомогательные ресурсы – любые другие ресурсы, не упомянутые ранее, которые могут быть использованы для решения поставленной задачи обработки языка. К ним можно отнести, например, корпусы параллельных текстов на целевом и некотором другом языке, автоматический переводчик на целевой язык, базы знаний, справочники и т. д.

Здесь и далее под языком может пониматься не только любой естественный или искусственный язык, но и некоторое подмножество естественного языка, например профессиональный жаргон моряков или русский язык опубликованных в Интернете текстов.

Обозначим основные ситуации, в которых можно столкнуться с дефицитом ресурсов [1]:

1. Целевой язык – малоизученный язык с малым числом носителей, например один из коренных языков жителей Южной Америки. Такому языку свойственен жесткий дефицит всех ресурсов.

2. Целевой язык – язык, число носителей которого относительно велико, но который не попадал во внимание компьютерных лингвистов. В качестве примера можно привести распространенные на полуострове Сомали кушитские языки: сидамо, хатия, камбата, каждым из которых владеет более миллиона человек. Для такого языка может наблюдаться недостаток оцифрованных текстов, корпусов размеченных и неразмеченных данных.

3. Целевой язык подробно исследован с точки зрения автоматической обработки (например, английский, русский), однако поставлена нестандартная задача или выбрана нетривиальная предметная область.

В настоящей работе исследован первый сценарий, касающийся обработки находящихся под угрозой исчезновения языков с малым числом носителей. В дальнейшем будем именовать их «малыми языками» по аналогии с расхожим термином «малые народы». Подчеркнем, что многие из изложенных ниже соображений будут применимы и к прочим сценариям, особенно к случаю крупных, но малоизученных языков.

Рассмотрим конкретный малый язык – язык южноамериканских индейцев туюка, относящийся к туканской языковой семье. Для оценки количества доступных ресурсов на этом языке был составлен корпус неразмеченных текстов. Анализ библиографической базы данных Glottolog[1], архива общественной организации SIL International[2] и базы языковых ресурсов Open Language Archives Community[3] показал, что в настоящее время существует менее 100 материалов, относящихся к языку туюка, причем некоторые из материалов представляют собой переиздание ранних публикаций. Многие из книг не были оцифрованы и недоступны в электронном виде. Важной составляющей корпуса неразмеченных текстов является перевод Нового Завета (URL: http://gospelgo.com/s/tuyuca_nt.htm). Вместе с ним в корпус вошли ряд опубликованных текстов на туюка, фразы из испанско-туюка и туюка-испанского словарей, а также предложения, использованные в качестве примеров в грамматических очерках. Суммарный объем корпуса составил около 180 000 слов. Для языковых моделей с современной архитектурой такого объема недостаточно, для их обучения используются корпуса на миллиард слов. Таким образом, моделирование языка туюка производится в условиях дефицита ресурсов.

2. Актуальность проблемы моделирования малых языков. В поддержку актуальности проблемы обработки языков с малым числом ресурсов можно высказать следующие соображения.

Во-первых, для носителей малого языка система обработки родного языка имеет такую же ценность, как для носителей английского, русского или любого другого крупного языка. Ценность представляют, например, механизмы рекомендаций в социальных сетях, способные анализировать интересы пользователя и предлагать ему другие публикации на схожие темы, автоматическим образом определяя их тему и тональность. Другой пример – интерфейсы взаимодействия с искусственными интеллектуальными системами на естественном языке, в частности чат-боты, способные поддержать диалог с пользователем, распознать некоторую типовую проблему и подсказать ее решение. Наконец, что, может быть, наиболее значимо, подобные системы могут стать шагом в направлении устранения проблемы информационного неравенства, от которого могут страдать носители редких и малоизученных языков. В настоящее время подавляющее большинство материалов в глобальной сети Интернет доступно на крупных языках, в первую очередь на английском, а также на русском, китайском, испанском и т. д. Эта информация включает в себя статьи и рекомендации по медицине, экономике, бытовым вопросам, разнообразные обучающие материалы и инструкции. Носители малых языков, не владеющие в достаточной степени крупными языками, не могут потреблять эту информацию. Отсюда возникает существенная социальная проблема информационного неравенства: не все жители Земли в равной мере обеспечены доступом к накопленной человечеством информа-

[1]Spoken L1 Language: Tuyuca [Electronic resource] / eds.: H. Hammarström, R. Forkel, M. Haspelmath ; Max Planck Institute for the Science of Human History // Glottolog. – Mode of access: https://glottolog.org/resource/languoid/id/tuyu1244. – Date of access: 13.10.2021.

[2]Language & Culture Archives: Tuyuca [Electronic resource] // SIL International. – Mode of access: https://www.sil.org/resources/search/language/tue. – Date of access: 13.10.2021.

[3]OLAC resources in and about the Tuyuca language [Electronic resource] // OLAC: Open Language Archives Community. – Mode of access: http://www.language-archives.org/language.php/tue. – Date of access: 13.10.2021.

ции. В некоторых случаях эта проблема может стать критической. Так, например, вызовом человечеству стала пандемия коронавируса, начавшаяся в 2020 г. Возникла экстренная необходимость проинформировать буквально каждого жители планеты о тех мерах предосторожности, которые следует предпринять в целях противостояния вирусу. Ответом на этот вызов стали решения, позволяющие автоматически переводить медицинские материалы на малые языки [3, 4]. В целом разработка эффективных автоматических переводчиков на малые языки является существенным шагом в направлении устранения информационного неравенства.

Во-вторых, ограничение на ресурсы мотивирует компьютерных лингвистов на поиск новых подходов к проблеме моделирования языков. В 2019 г. большое распространение получили модели, основанные на архитектуре нейронных сетей Transformer [5], например BERT [6]. Они оказались очень эффективными при решении разнообразных задач обработки естественного языка. Для подобных моделей применяется обучение с частичным привлечением учителя (semi-supervised learning) [2]: на первом этапе происходит обучение без учителя на большом корпусе неразмеченных данных, на втором – дообучение под конкретную задачу на небольшом количестве размеченных примеров, так называемый fine-tuning. Такой подход реализует идею переноса знаний (transfer learning): представление, полученное о языке в ходе первой стадии обучения, переиспользуется при решении конкретной задачи. Один из наиболее существенных недостатков указанного подхода заключается в том, что первая стадия обучения требует больших корпусов текстов. Например, в работах [7, 8] было показано, что корпус на несколько миллиардов слов не раскрывает весь потенциал модели BERT. Очевидно, что подобные корпуса недоступны для малых языков, не имеющих развитой литературной традиции и широкого представления в Интернете. Возникает отдельная задача моделирования малых языков, требующая собственных подходов. Наиболее популярное решение в этой области – адаптация имеющихся решений к работе в условиях дефицита ресурсов. Среди используемых подходов можно отметить две категории [1]:

– автоматическое расширение обучающей выборки (перенос аннотаций с текста на крупном языке на параллельный текст на моделируемом языке; автоматическое порождение данных, например, путем замены синонимов в имеющейся выборке);

– переиспользование знаний, полученных в ходе моделирования крупного языка (параллельное обучение модели многим языкам, сопоставление пространств векторных представлений слов на разных языках).

Обозначенные приемы имеют очевидные недостатки. Искусственная генерация похожих данных обедняет выборку. Совместное изучение нескольких языков эффективно, когда они родственны или типологически близки, но вызывает трудности при работе со своеобразным языком, например языком-изолятом. Представляет интерес разработка принципиально иных подходов, спроектированных специально для низкоресурсных языков.

В-третьих, автоматические модели могут сыграть важную роль в качестве инструмента сохранения языка, находящегося под угрозой исчезновения. Невозможно сохранить язык, на котором не желают разговаривать люди. Мотивация изучать малый язык имеет различную природу. Например, это может быть желание сохранить собственную идентичность и культуру. Существенным аргументом в пользу изучения языка может стать наличие доступных обучающих материалов, а также контента на разнообразные темы, в том числе художественных произведений. Соответствующие технологии позволяют упростить процесс изучения языка и облегчить создание текстов на целевом языке.

Актуальность проблемы сохранения малых языков, в свою очередь, неоднократно обсуждалась в литературе (см., например [9]). Можно отметить следующие соображения:

1. Для представителей соответствующих этносов родной язык может выступать одним из средств самоидентификации, возможно, основным, что становится особенно актуальным в эпоху глобализации. Более того, даже если в настоящее время общество отказывается от родного языка в пользу более крупного и развитого, со временем вновь может возникнуть спрос на культуру и язык предков.

2. Каждый язык является самоценным историко-культурным феноменом, в силу тесной связи языка и мышления являющимся своеобразным отражением уникальной философии народа,

сформировавшего этот язык, что, безусловно, значимо для таких наук, как этнография, этнология, культурная антропология.

3. Каждый язык служит дополнительным источником информации для лингвистических учений, причем каждая черта естественного языка дает дополнительную информацию об устройстве языка и речи вообще. Эти сведения используются в исследованиях человеческого мозга и мышления.

3. Освоение языка ребенком. Как было сказано выше, подходы к обработке малых языков в основном сводятся к адаптации идей, заложенных в построении крупных языковых моделей, подобных BERT, к языкам, располагающим малым числом доступных ресурсов для обучения. Речь идет либо об искусственном расширении обучающей выборки, либо о переносе знаний с модели крупного языка на модель низкоресурсного языка.

В этой связи любопытно задаться вопросом, как происходит освоение языка младенцем. Данный процесс универсален: вне зависимости от родного языка человек, не имеющий существенных особенностей в развитии, оказывается способен овладеть своим родным языком без каких-либо вспомогательных средств и методик [10, с. 2]. Можно ли в некотором приближении смоделировать этот процесс для обучения машины языку? Данный вопрос актуален не только с точки зрения психолингвистики, но и с точки зрения естественной обработки низкоресурсных языков: ребенок изучает родной язык, не имея каких-либо лингвистических сведений о нем, а также не опираясь на огромную обучающую выборку, т. е. использует тот объем ресурсов, который соизмерим с ресурсами, доступными для малоисследованного языка.

Вопрос языкового онтогенеза является открытым. Какие механизмы позволяют ребенку эффективно изучить родной язык, а также почему эти механизмы перестают функционировать после определенного возраста [10, с. 2], до сих пор неизвестно. Имеет смысл рассмотреть основные гипотезы, принятые в онтолингвистике, а также объективные факты, обнаруженные в ходе воспроизводимых экспериментов.

Стадии становления языка не раз описаны в литературе [11, с. 14–15; 12, с. 46–51; 13, с. 75–85]. Исследования показывают, что уже при рождении ребенок имеет определенную предрасположенность к языку, в частности способен узнавать по звучанию родной язык. Начиная с трех месяцев, у ребенка активируется функция запоминания слов, а в возрасте около пяти-семи месяцев ребенок начинает лепетать. В восемь месяцев у ребенка отдельные слова увязываются с объектами действительности – в первую очередь теми, которые движутся, т. е. выделяются на фоне неподвижной картинки. В возрасте около года человек начинает произносить полноценные слова, однако чаще всего они характеризуют ситуацию в целом и, возможно, его эмоциональное состояние. Такие слова получили название «голофразы». При игре с лошадкой высказанное «тпру» может обозначать и «лошадь», и «сани», и «садись», и «поедем», и «остановись». С возникновением представления о грамматике языка происходит разделение слов. На примере лошадки участники ситуации могут получить название «тпрунька», тогда как для действий будет выбрано другое слово. На более позднем этапе происходит выделение из контекста концепта лошадки, который связывается с ярлыком «лошадь». Когда размер словаря достигает приблизительно 100 слов, ребенок начинает их комбинировать. В возрасте около полутора лет у ребенка начинается интенсивное наращивание активного словарного запаса. Между 2 и 2,5 годами в речи увеличивается число используемых аффиксов и служебных слов. Примерно в три года происходит резкое становление грамматики: ребенок за короткий промежуток времени овладевает синтаксисом и морфологией языка в значительном объеме. Этот процесс осуществляется параллельно с пониманием маленьким человеком иерархической структуры мира вообще: как вещи состоят из деталей, так и речь состоит из отдельных слов и компонентов. На этом этапе ребенок формулирует для себя определенные правила языка, которые последовательно уточняются путем проб и ошибок.

Безусловно, дети не рождаются со знанием языка и навык владения языком не вырабатывается автоматически. Что именно является движущей силой, заставляющей ребенка учить язык, достоверно неизвестно. На этот счет существуют различные теории.

Бихевиористическая теория научения предполагает, что основными механизмами освоения языка являются подражание и подкрепление. Ребенок пробует подражать речи своих опекунов

и ориентируется на их реакцию («подкрепляя» ту или иную гипотезу). В целом данная гипотеза не позволяет полностью объяснить процесс становления речи в полной мере [14, с. 74–75].

Противоположная теория, базирующаяся на идеях Н. Хомского, заключается в том, что ребенок рождается сразу с сильными предпосылками к изучению языка: в его мозге наличествует специальная программа, которая обеспечивает усвоение грамматики любого языка, – так называемая универсальная грамматика. Ориентируясь на звучащую речь, ребенок адаптирует правила универсальной грамматики к своему родному языку, осуществляя тем самым обучение [14, с. 75–77]. Данная теория также подвергается критике: опыт детей, лишенных возможности изучать язык (выросших в лесу или подвергнутых жестокому обращению со стороны опекунов), показывает, что у них универсальная грамматика не реализуется никаким образом, хотя при восприятии языка как инстинкта более естественно было бы ожидать случайной реализации [15, с. 100]. Кроме того, в универсальной грамматике должны были быть некоторые универсальные элементы, свойственные всем языкам. Исследования показывают, что обнаружить подобные универсалии не представляется возможным [15, с. 93–94].

Когнитивная теория происхождения речи человека предполагает, что развитие речи обусловлено присущей ребенку с рождения способностью получать и обрабатывать информацию. В отличие от теории научения когнитивная теория предполагает, что движущим механизмом изучения языка является не подражание, а социальное взаимодействие [15, с. 74–75]. Согласно этой теории развитие языка не отличается от развития восприятия, памяти или мыслительных процессов.

Ребенок учится участвовать в разговоре, больше узнавая о языке и о том, как им пользоваться, тренируется планировать разговор на языке [11, с. 5–6]. Участие взрослого в изучении языка ребенком очень важно. Эксперименты показывают, что ребенок не в состоянии выучить язык, регулярно слушая телевизор. Взрослый, участвуя в разговоре с ребенком, следит за проявлениями его внимания и имеет возможность в зависимости от обстоятельств корректировать свое речевое поведение. Кроме того, речь взрослых при общении с ребенком имеет целый ряд специфических черт, призванных способствовать овладению языком: она медленнее, в ней строже соблюдаются правила грамматики, она больше нацелена на происходящее здесь и сейчас, в ней встречается больше повторов [13, с. 79].

4. Освоение языка и обучение с подкреплением. Изучение языка не происходит в отрыве от изучения мира – это два взаимно обусловленных процесса. Советский психолог Л. С. Выготский сближал факт развития значения слова с фактом развития сознания. Для него слово – это средство, которое отражает внешний мир в его связях и отношениях [12, с. 42]. На вышеуказанном материале наблюдаются параллели между познанием мира и изучением языка: ребенок одновременно приходит к пониманию иерархичности мира и иерархичности языка [13, с. 82].

Тесная связь между языком и мышлением, между изучением языка и изучением окружающего мира выводит задачу моделирования онтогенеза языка за рамки компьютерной лингвистики и приближает ее к задаче моделирования искусственного интеллекта.

Если рассматривать язык и мир с точки зрения семиотики Чарльза Пирса, то эти два сложных компонента могут быть представлены двумя моделями: моделью языка как многоуровневой системы знаков и моделью мира как системы произвольного вида. Процесс изучения в этом случае будет сводиться к изучению правил функционирования обеих моделей, а также выработке соотношения между компонентами двух моделей – «значениями» [16, с. 76].

Подобной структурой обладала, например, «программа, понимающая естественный язык» Терри Винограда [17]. Предложенный Виноградом агент SHRDLU действовал в «мире», состоящем из блоков различных форм и цветов, выполнял инструкции по переносу блоков, сохранял историю инструкций, мог изучать и оперировать новыми понятиями. Стоит отметить, что существенный пласт лексики был задан аксиоматически, как такового изучения естественного языка решение не предполагало. Проблему сопоставления информации, получаемой по двум каналам, один из которых – текст на естественном языке, затрагивает задача Visual Question Answering [18] и ее развитие – задача Embodied Question Answering [19]. Задача Visual Question Answering предполагает разработку программы, способной отвечать на вопросы, сформулированные на естественном языке и относящиеся к изображению, также поступающему на вход

системы: «Сколько лошадей видно?» – «Две лошади». В задаче Embodied Question Answering агент действует внутри своеобразного 3D-мира и для поиска ответа ему требуется совершать дополнительные действия (в частности, перемещаться в пространстве, чтобы увидеть предмет, к которому относится вопрос).

И бихевиористическая, и когнитивная гипотезы усвоения языка предполагают, что ребенок, пытаясь использовать речь, преследует некоторые цели (подражание или социальное взаимодействие). Эффективность коммуникации некоторым образом подкрепляется: была достигнута цель или нет. Этот процесс во многом напоминает процесс обучения с подкреплением – одно из направлений машинного обучения.

Подробный обзор работ, находящихся на стыке обучения с подкреплением и обработки естественного языка, рассмотрен в статье [20]. Многие из исследований затрагивают задачу следования инструкциям, представленным на некотором языке. Например, агент, разработанный в публикации [21], учится ориентироваться в 2D-мирах определенного вида и перемещаться в точку, заданную с помощью команды на английском языке (Reach the cell above the westernmost rock). Текстовые квесты представляют особый интерес, так как в этом случае на естественном языке показаны не только инструкции, но и текущее состояние окружения. Для построения подобных текстовых миров был разработан генератор TextWorld [22].

Необходимо отметить, что эффективность агента в отмеченных работах определяется по его способности достигать цель, а не по уровню овладения языком. Если ставить целью освоение естественного языка, следует продумать функцию награды как неотъемлемого компонента обучения с подкреплением. Эта функция может быть построена с помощью обратного обучения с подкреплением – метода, при котором заданы стратегии действия, рассматриваемые как «экспертные», «оптимальные», а цель обучения – построение корректной функции награды, которая в некотором роде объяснит действия «эксперта» [23].

Обучение с подкреплением продемонстрировало хорошие результаты при моделировании автоматических игроков в настольные игры. Например, нейронная сеть AlphaZero, разработанная компанией DeepMind, обучилась побеждать гроссмейстеров в шахматы, сеги и го [24]. При некотором допущении коммуникацию тоже можно рассматривать как игру, цель которой – получить от собеседника ожидаемый отклик. Текущий диалог может быть рассмотрен как состояние среды, возможные реплики агента – как потенциальные действия, а соответствие реплики собеседника ожидаемой – как функция оценки. Эта конфигурация соответствует теоретическому описанию обучения с подкреплением. Таким образом сформулирована задача, в которой агент методом проб и ошибок ищет стратегию оптимального использования естественного языка, что определенным образом перекликается с тем, как овладевает речью ребенок.

5. Существующие модели освоения языка. Вопрос моделирования онтогенеза языка исследовался в литературе. Как правило, целью таких исследований преимущественно является проверка психолингвистических гипотез о процессе изучения языка [25, с. 92].

Если говорить о процессе освоения словаря, то используемые подходы можно разбить на две категории. В первую категорию попадает изучение связи между словами и референтами, во вторую – постепенное уточнение смысла слов на базе полных предложений и некоторого, зачастую зашумленного, визуального представления [25, с. 94]. Второй подход применен, например, в работе [26], где в качестве «фона» используется множество меток, а результат обучения системы заключается в распределении вероятностей, что данное слово соотносится с данной меткой. Отдельного внимания удостаиваются синонимы и омонимы, которые нарушают отношение «один к одному». Авторы делают вывод, что для изучения смыслов слов в таком окружении не требуется особых предпосылок и специальных механизмов изучения, достаточно общих алгоритмов.

Изучение морфологии языка и обнаружение закономерностей словоизменения представляют собой задачи, которые могут быть решены с помощью нейронной сети прямого распространения [27]. Характер обучения соотносится с наблюдаемым процессом изучения морфологии детьми – так называемой U-образной кривой обучения: вначале ребенок запоминает конкретные формы, затем совершает обобщение, которое может приводить к ошибкам (например,

форма Past Simple в английском языке обычно образуется с помощью суффикса -ed, однако для ряда неправильных глаголов это утверждение неверно), и, наконец, выучивает исключения.

Моделирование освоения грамматики является более сложной задачей. В терминах теории универсальной грамматики решить ее можно только с помощью моделирования параметризованной грамматики, способной реализоваться в виде различных естественных языков. Публикация [25] – один из примеров такой модели: в ее основе лежит алгоритм, способный изучить категориальную грамматику.

Если же исходить из бихевиористической или когнитивной теорий, грамматику можно «извлечь» из конкретных примеров с помощью общих алгоритмов. Были разработаны модели, которые показали, что с помощью статистического анализа из массива высказываний, адресованных ребенку, можно извлечь информацию о структурах и закономерностях языка [25, с. 94; 28, с. 48]. В качестве материала исследования используются корпуса детской речи, например база CHILDES [29], содержащая коллекцию экспериментальных и наблюдательных данных об изучении языка детьми, а также о взаимодействии детей и взрослых.

Интерес представляет, например, модель MOSAIC (Model of Syntax Acquisition in Children) [30], обучающаяся на неразмеченном тексте и способная порождать высказывания, похожие на детские. Модель MOSAIC обучается в два этапа: на первом этапе строится иерархическая структура связанных ячеек, где каждая ячейка символизирует слово, а каждая связь представляет собой разницу между связанными словами. На втором этапе порождаются новые связи между словами со схожими контекстами. Этот шаг позволяет обнаруживать в речи закономерности и делать обобщения [31, с. 52]. Отмечается, что после обучения системы на корпусе CHILDES она достаточно эффективно имитировала особенности детской речи.

6. Архитектура модели. Безусловно, имитация онтогенеза осложнена тем, что само это явление изучено недостаточно. В литературе представлены опыты, имитирующие отдельные аспекты изучения языка ребенком. Обучающая выборка в этих экспериментах строится с ориентиром на объем реплик, доступных ребенку к некоторому возрасту, а результаты, в свою очередь, соотносятся с тем, насколько хорошо дети владеют речью. Таким образом удается сымитировать изучение лексики, словоизменения и грамматики.

Построение системы имитации онтогенеза – очень трудоемкая задача, особенно с учетом обозначенной выше взаимосвязи между изучением языка и изучением мира. Прежде всего следует определить те аспекты этого процесса, которыми можно пренебречь. В частности, педагоги отмечают, что процесс изучения жестового языка у глухих детей идентичен изучению звукового языка у слышащих детей. В частности, дети родителей, использующих жестовый язык, лепечут руками [13, с. 76]. Известно, что неслышащие люди, у которых повреждена зона Брока, имеют проблемы с высказываниями на языке жестов – точно так же, как люди, использующие звуковой язык, испытывают значительные затруднения на морфологическом и синтаксическом уровнях [13, с. 51]. У слабовидящих и слепых детей усвоение речи также осуществляется в процессе общения и по тем же закономерностям, что и у зрячих, хотя и с особенностями, вызванными объективными причинами [32]. Это позволяет предположить, что природа каналов связи с внешним миром не является существенной. Следовательно, при изучении принципиальной эффективности предлагаемого подхода допустимо выбрать произвольную модальность речи. В дальнейшем в настоящей работе будем ориентироваться на текстовое представление языка. Предполагается, что если предложенная система окажется эффективной, ее можно будет адаптировать для работы со звуковым представлением языка без существенных изменений в ядре архитектуры: либо непосредственно с помощью подготовленной обучающей выборки звуковых материалов, либо посредством обучаемого компонента, конвертирующего произнесенные фразы в текст.

В полноценную систему, результат работы которой будет использоваться как языковая модель, можно включить перечисленные выше разработки. Отметим, что они требуют значительной доработки. Так, например, в эксперименте с изучением связей между словами и «объектами» между двумя множествами строилось биективное отображение. Доступное ребенку предположение, что каждый объект имеет какое-то свое название, нередко рассматривается в психолингвистической литературе как один из двигателей процесса усвоения языка. Обрат-

ное предположение в корне неверно: многие слова выражают значительно более сложные концепции – от плохо формализуемых абстрактных понятий вроде цвета, запаха или ощущения до логических операторов, например кванторных местоимений «все», «каждый». Следует предусмотреть возможность изучения названий характеристик, численных мер или действий.

Модель нужно дополнить вспомогательными компонентами, в первую очередь имитацией «мира», объекты которого будут соотноситься со словами. В описанном выше эксперименте с изучением лексики «мир» (или «сцена» в терминах публикации) представлял собой неупорядоченный набор меток. В этом случае, однако, связи между объектами примитивны. Выбор более сложной конфигурации «мира» (например, блочного мира Терри Винограда или шахматной доски) позволит также моделировать лексику, соответствующую положению объектов, их атрибутов и взаимодействию между собой. Между тем это усложняет эксперимент, поскольку требует выработки схемы представления упомянутых концепций.

Другой модуль, который может быть добавлен, – компонент, оценивающий эффективность коммуникации. Как отмечают психологи, желание успешно взаимодействовать с окружающими также является одним из факторов, обусловливающих эффективное освоение языка ребенком. Этот фактор может некоторым образом порождать стремление совершить коммуникативный акт, а затем сопоставить воспринятый ответ с ожиданиями и соответствующим образом оценить состоявшийся диалог. Полученная оценка может быть использована как функция награды в терминах обучения с подкреплением.

Можно предложить и другие дополнительные компоненты, в частности средство поддержания контекста, позволяющее поддерживать тему диалога на протяжении нескольких реплик, или базу знаний, хранящую освоенную ранее информацию.

Попробуем соединить все описанные концепты воедино. Инициатором коммуникации выступает своеобразный «мозг» системы. Этот модуль в первую очередь имеет доступ ко всем компонентам, представляющим собой источник какой-либо информации: «миру», или «сцене», средству сохранения контекста, базе знаний. Выше упоминалось о взаимосвязи изучения языка и познания мира ребенком, однако, безусловно, следует понимать, что разработка модели «познания мира» слишком сложна. На этом этапе необходимо внести ряд упрощений. Рассматриваемый модуль можно представить как чат-бот-систему, которая принимает на вход реплику собеседника (в том числе нулевую, что будет соответствовать самостоятельной инициации беседы) и данные «окружающего мира», а затем генерирует реплику-ответ. Поскольку имитации сознания не требуется, можно воспользоваться одним из готовых алгоритмов, в частности нейронной сетью на базе архитектуры Transformer [33].

Появление нейронной сети сложной архитектуры вновь требует обучения на огромном массиве данных, недоступном для малого языка. Обойти эту проблему предполагается с помощью промежуточного метаязыка. В данном случае реплики будут проходить три стадии: реплика на исходном языке, представление реплики на метаязыке и векторное представление реплики с помощью Transformer. Аналогично идет обратный процесс: векторное представление реплики порождает представление на метаязыке, которое в свою очередь становится основой реплики на естественном языке. Transformer может быть использован как чат-бот, оперирующий метаязыком. Такого чат-бота можно обучить на материалах крупного языка (английского или русского) с помощью промежуточного слоя, ответственного за перевод с естественного языка на метаязык и обратно.

Структура метаязыка – предмет дальнейшего изучения. В целом он должен позволять выразить основные компоненты высказывания, такие как субъект, объект, предикат, атрибут. Подобные вопросы формального представления смысла проработаны в рамках формальной семантики [34].

Интерес вызывает представление лексики. В работе [26] всем словам, используемым в эксперименте, соответствовали метки на «сцене». В предлагаемой модели количество объектов «мира» (блочного мира Терри Винограда, шахматной доски) ограничено. Ожидается, что для лексики, соответствующей объектам «мира», будет найдено предметное соответствие.

Что касается остальных слов, можно рассмотреть следующий подход. По умолчанию предполагается, что неизвестным словам будут соответствовать некоторые нумерованные невиди-

мые объекты. Отношения между словами будут изучаться на примерах их употребления. В первую очередь речь идет о таком соотношении, как тождественность и различие. Более сложный подход предполагает изучение смыслов слов как результат применения логических функций к другим ранее изученным словам. Поиск подобного семантического представления для лексем широко представлен в работах А. Вежбицка и К. Годдард: авторы стремятся построить метаязык, который позволит определить все лексемы языка посредством малого набора «семантических примитивов» [35]. Важно, чтобы система имела возможность изучать не только существительные. Поэтому следует продумать способ представления «мира». Если хранить шахматную доску как двойной массив, элементы которого – фигуры, у системы не будет возможности изучить понятие «находится». Вместо этого следует использовать предикаты с аргументами, например «Находится(Конь, Доска)», «Соседствует(Конь, Слон)», «ОбладаетКачеством(Конь, белый)».

Следует понимать, что при переводе с метаязыка на естественный язык могут возникнуть непредвиденные трудности. Например, во многих языках коренных народов Южной Америки грамматикализирована эвиденциальность – указание на источник информации. В языке туюка туканской языковой семьи (его носители проживают на границе Бразилии и Колумбии) каждый глагол обязательно получает один из пяти суффиксов: визуального свидетельства (díiga apéwi – он играл в футбол, говорящий это видел), невизуального свидетельства (díiga apéti – он играл в футбол, говорящий слышал, как это происходило), вещественного свидетельства (díiga apéji – он играл в футбол, и говорящий обнаружил тому явные подтверждения), пересказывательности (díiga apéjigɨ – он играл в футбол, и говорящему об этом кто-то сообщил) и умозаключения (díiga apéhĩji – он играл в футбол, ибо у говорящего есть причины так считать) [36]. Если метаязык будет проектироваться на основе европейских языков, он не будет обязательно хранить значение эвиденциальности. Как результат, сформулированное на метаязыке высказывание «он играл в футбол», вероятнее всего, не получится корректно перевести на туюка: придется искусственно делать предположение об источнике информации и выбор суффикса окажется произвольным.

Конвертация метаязыка в малый естественный язык и обратно должна осуществляться с помощью компонентов, описанных в разд. 5. Грамматический модуль обучается строить структуру фразы на языке, лексический модуль тренируется подбирать подходящие лексемы, морфологический модуль совершенствуется в подборе верной словоформы. Обучение компонентов может происходить как на основе успешности коммуникации, так и по прецедентам с использованием в качестве обучающей выборки ответных реплик.

Заключение. В работе подробно обоснована актуальность моделирования малых языков, показана социальная значимость проблемы, польза ее решения для лингвистики, этнографии, этнологии и культурной антропологии, отмечено потенциальное развитие сферы обработки естественного языка.

В отличие от крупных языков, эффективно моделируемых с помощью решений на базе Transformer, малые языки не обладают достаточным количеством доступных ресурсов и потому не могут быть обработаны столь же успешно. В статье рассмотрена возможность моделирования малых языков с помощью имитации онтогенеза. Для этого приведен краткий обзор современного научного представления об онтогенезе языка. В полной мере данное явление и его движущие механизмы не изучены. Тем не менее накопленные сведения позволяют моделировать отдельные аспекты процесса освоения языка, в том числе опираясь на психолингвистические гипотезы. Можно говорить о связанном развитии психолингвистики и языкового моделирования: гипотезы о ходе процесса усвоения языка позволяют проектировать те или иные модели, а эффективность моделей в свою очередь позволяет характеризовать правдоподобность гипотез. Определенные результаты в этой сфере представлены в научной литературе: были разработаны системы, которые изучали лексический, морфологический и грамматический уровни языка, опираясь на корпус детской речи CHILDES и во многом имитируя усвоение соответствующих аспектов языка ребенком.

В контексте проблемы моделирования познания языка, которая сопряжена с познанием мира, можно вспомнить алгоритмы машинного обучения с подкреплением, мотивированные пси-

хическими моделями. Основная тенденция в современных работах, находящихся на стыке обработки языка и обучения с подкреплением, – обучение агента исполнению команд, отданных на естественном языке. Изучение языка в этих экспериментах является вспомогательной задачей. Однако успешную коммуникацию можно рассматривать и как цель, к которой стремится агент. В этом случае система обучения с подкреплением может преследовать цель изучения языка как основную.

На основании приведенных соображений в работе представлен концепт системы обработки языка, обучение которой происходит посредством моделирования онтогенеза. Выделены основные компоненты системы и принципы их взаимодействия.

Список использованных источников

1. A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios [Electronic resource] / M. A. Hedderich [et al.]. – 2020. – Mode of access: https://arxiv.org/abs/2010.12309. – Date of access: 12.10.2021.

2. Dai, A. M. Semi-supervised sequence learning [Electronic resource] / A. M. Dai, Q. V. Le // Proc. of the 28th Intern. Conf. on Neural Information Processing Systems. – 2015. – Vol. 2. – P. 3079–3087. https://doi.org/10.18653/v1/P17-1161

3. TICO-19: the translation initiative for Covid-19 [Electronic resource] / A. Anastasopoulos [et al.] // Proc. of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020. – Dec. 2020. – Mode of access: https://aclanthology.org/2020.nlpcovid19-2.5/. – Date of access: 12.10.2021. https://doi.org/10.18653/v1/2020.nlpcovid19-2.5

4. Enabling low-resource transfer learning across Covid-19 corpora by combining event-extraction and co-training / A. Spangher [et al.] // Proc. of the 1st Workshop on NLP for COVID-19 at ACL 2020. – July 2020. – Mode of access: https://aclanthology.org/2020.nlpcovid19-acl.4/. – Date of access: 12.10.2021.

5. Attention is all you need / A. Vaswani [et al.] // Proc. of the 31st Intern. Conf. on Neural Information Processing Systems, Long Beach, California, USA, 4–9 Dec. 2017. – Long Beach, 2017. – P. 6000–6010.

6. Качков, Д. И. Моделирование языка и двунаправленные представления кодировщиков: обзор ключевых технологий / Д. И. Качков // Информатика. – 2020. – Т. 17, № 4. – С. 61–72. https://doi.org/10.37661/1816-0301-2020-17-4-61-72

7. Cloze-driven pretraining of self-attention networks / A. Baevski [et al.] // Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Intern. Joint Conf. on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 Nov. 2019. – Hong Kong, 2019. – P. 5360–5369. https://doi.org/10.18653/v1/D19-1539

8. RoBERTa: A Robustly Optimized BERT Pretraining Approach [Electronic resource] / Y. Liu [et al.]. – 2019. – Mode of access: https://arxiv.org/abs/1907.11692. – Date of access: 12.10.2021.

9. Замятин, К. Как и зачем сохранять языки народов России / К. Замятин, А. Пасанен, Я. Саарикиви. – Хельсинки, 2012. – 181 с.

10. Meisel, J. M. First and Second Language Acquisition (Cambridge Textbooks in Linguistics) / J. M. Meisel. – Cambridge University Press, 2011. – 318 p.

11. Clark, E. V. First Language Acquisition / E. V. Clark. – Cambridge University Press, 2009. – 2nd ed. – 490 p.

12. Лурия, А. Р. Язык и сознание / А. Р. Лурия ; под ред. Е. Д. Хомской. – М. : Изд-во Моск. ун-та, 1979. – 320 с.

13. Бурлак, С. А. Происхождение языка. Факты, исследования, гипотезы / С. А. Бурлак. – М. : Альпина Диджитал, 2019. – 609 с.

14. Немов, Р. С. Общая психология в 3 т. Том II в 4 кн. Книга 4. Речь. Психические состояния : учебник и практикум для академического бакалавриата / Р. С. Немов. – 6-е изд., перераб. и доп. – М. : Юрайт, 2017. – 243 с.

15. Evans, V. The Language Myth Why Language Is Not an Instinct / V. Evans. – Cambridge University Press, 2014. – 314 p.

16. Пирс, Ч. С. Принципы философии : в 2 т. / Ч. С. Пирс ; пер. с англ. В. В. Кирющенко, М. В. Колопотина. – СПб. : Санкт-Петербургское философское общество, 2001. – Т. 2. – 313 с.

17. Виноград, Т. Программа, понимающая естественный язык / Т. Виноград. – М. : Мир, 1976. – 296 с.

18. VQA: visual question answering / S. Antol [et al.] // IEEE Intern. Conf. on Computer Vision (ICCV). – Santiago, Chile, 2015. – P. 2425–2433. https://doi.org/10.1109/ICCV.2015.279

19. Embodied question answering / A. Das [et al.] // Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018. – Salt Lake City, 2018. – P. 1–10.

20. A survey of reinforcement learning informed by natural language / J. Luketina [et al.] // Proc. of the Twenty-Eighth Intern. Joint Conf. on Artificial Intelligence, Macao, China, 10–16 Aug. 2019. – Macao, 2019. – P. 6309–6317. https://doi.org/10.24963/ ijcai.2019/880

21. Janner, M. Representation learning for grounded spatial reasoning / M. Janner, K. Narasimhan, R. Barzilay // Transactions of the Association for Computational Linguistics. – 2018. – Vol. 6. – P. 49–61. https://doi.org/10.1162/tacl_a_00004

22. Côté, M.-A. TextWorld: A learning environment for text-based games / M.-A. Côté ; T. Cazenave, A. Saffidine, N. Sturtevant (eds.) // Computer Games. CGW 2018. Communications in Computer and Information Science. – Cham : Springer, 2018. – Vol. 1017. – P. 41–75. https://doi.org/10.1007/978-3-030-24337-1_3

23. Arora, S. A survey of inverse reinforcement learning: Challenges, methods and progress [Electronic resource] / S. Arora, P. Doshi // Artificial Intelligence. – 2021. – Vol. 297. – Mode of access: https://arxiv.org/abs/1806.06877. – Date of access: 12.10.2021. https://doi.org/10.1016/j.artint.2021.103500

24. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play / D. Silver [et al.] // Science. – 2018. – Vol. 362, no. 6419. – P. 1140–1144. https://dx.doi.org/10.1126%2Fscience.aar6404

25. Freudenthal, D. Computational models of language development / D. Freudenthal, A. Alishahi ; P. J. Brooks, V. Kempe (eds.) // Encyclopedia of Language Development. – 1st ed. – SAGE Publications Inc., 2014. – P. 92–96.

26. Fazly, A. A probabilistic computational model of cross-situational word learning / A. Fazly, A. Alishahi, S. Stevenson // Cognitive Science. – 2010. – Vol. 34, iss. 6. – P. 1017–1063. https://doi.org/10.1111/j.1551-6709.2010.01104.x

27. Christiansen, M. H. Connectionist natural language processing: the state of the art / M. H. Christiansen, N. Chater // Cognitive Science. – 1999. – Vol. 23, iss. 4. – P. 417–437. https://doi.org/10.1207/s15516709cog2304_2

28. Buttery, P. J. Computational models for first language acquisition / P. J. Buttery // Technical Report UCAM-CL-TR-675. – University of Cambridge, 2006. – Mode of access: https://www.cl. cam.ac.uk/techreports/UCAM-CL-TR-675.pdf. – Date of access: 21.03.2021.

29. MacWhinney, B. The CHILDES Project: Tools for Analyzing Talk: Transcription Format and Programs (3rd ed.) / B. MacWhinney. – Lawrence Erlbaum Associates Publishers, 2000.

30. Jones, G. A process model of children's early verb use / G. Jones, F. Gobet, J. M. Pine // Proc. of the 22th Annual Conf. of the Cognitive Science Society, Philadelphia, PA, 13–15 Aug. 2000. – Philadelphia, 2000. – P. 723–728.

31. Alishahi, A. Computational Modeling of Human Language Acquisition / A. Alishahi. – Morgan & Claypool, 2010. – 107 p.

32. Andersen, E. S. The impact of input: language acquisition in the visually impaired / E. S. Andersen, A. Dunlea, L. Kekelis // First Language. – 1993. – Vol. 13, no. 37. – P. 23–49. https://doi.org/10.1177/014272379301303703

33. Vlasov, V. Dialogue Transformers [Electronic resource] / V. Vlasov, J. E. M. Mosig, A. Nicho. – 2019. – Mode of access: https://arxiv.org/abs/1910.00486. – Date of access: 12.10.2021.

34. Андреев, А. В. Введение в формальную семантику : учеб. пособие / А. В. Андреев, О. А. Митрофанова, К. В. Соколов. – СПб. : СПбГУ, 2014. – 88 с.

35. Goddard, C. The search for the shared semantic core of all languages / C. Goddard ; C. Goddard, A. Wierzbicka (eds.) // Meaning and Universal Grammar – Theory and Empirical Findings. – Amsterdam : John Benjamins, 2002. – Vol. I. – P. 5–40.

36. Barnes, J. Evidentials in the Tuyuca Verb / J. Barnes // Intern. J. of American Linguistics. – 1984. – Vol. 50, no. 3. – P. 255–271.

References

1. Hedderich M. A., Lange L., Adel H., Strötgen J., Klakow D. *A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios*, 2020. Available at: https://arxiv.org/abs/2010.12309 (accessed 12.10.2021).

2. Dai A. M., Le Q. V. Semi-supervised sequence learning. *Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2015, vol. 2, pp. 3079–3087. https://doi.org/10.18653/v1/P17-1161

3. Anastasopoulos A., Cattelan A., Dou Z.-Y., Federico M., Federman C., …, Tur S. TICO-19: the translation initiative for Covid-19. *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. December 2020. Available at: https://aclanthology.org/2020.nlpcovid19-2.5/ (accessed 12.10.2021). https://doi.org/10.18653/v1/2020.nlpcovid19-2.5

4. Spangher A., Peng N., May J., Ferrara E. Enabling low-resource transfer learning across Covid-19 corpora by combining event-extraction and co-training. *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. July 2020. Available at: https://aclanthology.org/2020.nlpcovid19-acl.4/ (accessed 12.10.2021).

5. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., …, Polosukhin I. Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, 4–9 December 2017*. Long Beach, 2017, pp. 6000–6010.

6. Kachkou D. I. Language modeling and bidirectional coders representations: an overview of key technologies. Informatika [*Informatics*], 2020, vol. 17, no. 4, pp. 61−72 (In Russ.). https://doi.org/10.37661/1816-0301-2020-17-4-61-72

7. Baevski A., Edunov S., Liu Y., Zettlemoyer L., Auli M. Cloze-driven pretraining of self-attention networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019.* Hong Kong, 2019, pp. 5360–5369. https://doi.org/10.18653/v1/D19-1539

8. Liu Y., Ott M., Goyal N., Du J., Joshi M., …, Stoyanov V. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*, 2019. Available at: https://arxiv.org/abs/1907.11692 (accessed 12.10.2021).

9. Zamyatin K., Pasanen A., Saarikivi Ya. Kak i zachem sohranyat' yazyki narodov Rossii. *How and Why to Save Languages of Ethnic Groups in Russia*. Helsinkim, 2012, 181 p. (In Russ.).

10. Meisel J. M. *First and Second Language Acquisition (Cambridge Textbooks in Linguistics)*. Cambridge University Press, 2011, 318 p.

11. Clark E. V. *First Language Acquisition*. Cambridge University Press, 2nd ed., 2009, 490 p.

12. Luriya A. R. Yazyk i soznanie. *Language and Conscience*. In Homskaya E. D. (ed.). Moscow, Izdatel'stvo Moskovskogo universtiteta, 1979, 320 p. (In Russ.).

13. Burlak S. A. Proishozhdenie yazyka. Fakty, issledovaniya, gipotezy. *Origin of the language. Facts, Researches, Hypothesis*. Moscow, Alpina digital, 2019, 609 p. (In Russ.).

14. Nemov R. S. Obshchaya psihologiya. *General Psychology.* Vol. 2, kniga 4. Rech'. Psihicheskie sostoyaniya: uchebnik i praktikum dlya akademicheskogo bakalavriata. *Speech. Psychological States: Textbook and Workshop for Bachelors*. 6th ed., Moscow, Yurait, 2017, 243 p. (In Russ.).

15. Evans V. *The Language Myth Why Language Is Not an Instinct*. Cambridge University Press, 2014, 314 p.

16. Peirce C. S. *Collected Papers of Charles Sanders Peirce, Volumes I and II: Principles of Philosophy and Elements of Logic*. Belknap Press, 1932, vol. II, 535 p.

17. Winograd T. Programma, ponimaushchaya estestvennyj yazyk. *Understanding Natural Language*. Moscow, Mir, 1976, 296 p. (In Russ.).

18. Antol S., Agrawal A., Lu J., Mitchell M., Batra D., …, Parikh D. VQA: visual question answering. *IEEE International Conference on Computer Vision (ICCV)*. Santiago, Chile, 2015, pp. 2425–2433. https://doi.org/10.1109/ICCV.2015.279

19. Das A., Datta S., Gkioxari G., Lee S., Parikh D., Batra D. Embodied question answering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.* Salt Lake City, 2018, pp. 1–10.

20. Luketina J., Nardelli N., Farquhar G., Foerster J., Andreas J., …, Rocktäschel T. A survey of reinforcement learning informed by natural language. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, Macao, China, 10–16 August 2019*. Macao, 2019, pp. 6309–6317. https://doi.org/10.24963/ijcai.2019/880

21. Janner M., Narasimhan K., Barzilay R. Representation learning for grounded spatial reasoning. *Transactions of the Association for Computational Linguistics*, 2018, vol. 6, pp. 49–61. https://doi.org/10.1162/tacl_a_00004

22. Côté M.-A., Kádár Á., Yuan X., Kybartas B., Barnes T., …, Trischler A. TextWorld: A learning environment for text-based games. *Computer Games. CGW 2018. Communications in Computer and Information Science*, Cham, Springer, 2018, vol. 1017, pp. 41–75. https://doi.org/10.1007/978-3-030-24337-1_3

23. Arora S., Doshi P. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, 2021, vol. 297. Available at: https://arxiv.org/abs/1806.06877 (accessed 12.10.2021). https://doi.org/10.1016/j.artint.2021.103500

24. Silver D., Hubert T., Schrittwieser J., Antonoglou I., Lai M., …, Hassabis D. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 2018, vol. 362, no. 6419, pp. 1140–1144. https://dx.doi.org/10.1126%2Fscience.aar6404

25. Freudenthal D., Alishahi A. Computational models of language development. *Encyclopedia of Language Development*. In Brooks P. J., Kempe V. (eds.). 1st ed., SAGE Publications Inc., 2014, pp. 92–96.

26. Fazly A., Alishahi A., Stevenson S. A probabilistic computational model of cross-situational word learning. *Cognitive Science*, 2010, vol. 34, iss. 6, pp. 1017–1063. https://doi.org/10.1111/j.1551-6709.2010.01104.x

27. Christiansen M. H., Chater N. Connectionist natural language processing: the state of the art. *Cognitive Science*, 1999, vol. 23, iss. 4, pp. 417–437. https://doi.org/10.1207/s15516709cog2304_2

28. Buttery P. J. Computational models for first language acquisition. *Technical Report UCAM-CL-TR-675*, University of Cambridge, 2006. Available at: https://www.cl.cam.ac.uk/techreports/ UCAM-CL-TR-675.pdf (accessed 21.03.2021).

29. MacWhinney, B. *The CHILDES Project: Tools for Analyzing Talk: Transcription Format and Programs*. 3rd ed., Lawrence Erlbaum Associates Publishers, 2000.

30. Jones G., Gobet F., Pine J. M. A process model of children's early verb use. *Proceedings of the 22th Annual Conference of the Cognitive Science Society, Philadelphia, PA, 13–15 August 2000*. Philadelphia, 2000, pp. 723–728.

31. Alishahi A. *Computational Modeling of Human Language Acquisition*. Morgan & Claypool, 2010, 107 p.

32. Andersen E. S, Dunlea A., Kekelis L. The impact of input: language acquisition in the visually impaired. *First Language*, 1993, vol. 13, no. 37, pp. 23–49. https://doi.org/10.1177/014272379301303703

33. Vlasov V., Mosig J. E. M., Nicho A. *Dialogue Transformers*, 2019. Available at: https://arxiv.org/abs/1910.00486 (accessed 12.10.2021).

34. Andreev A. V., Mitrofanova O. A., Sokolov K. V. Vvedenie v formal'nuyu semantiky: uchebnoe posobie. *Introduction Into Formal Semantics: Handbook*. Saint-Petersburg, Saint-Petersburg State University, 2014, 88 p. (In Russ.).

35. Goddard C. The search for the shared semantic core of all languages. In Goddard C., Wierzbicka A. (eds.). *Meaning and Universal Grammar – Theory and Empirical Findings*. Amsterdam, John Benjamins, 2002, vol. I, pp. 5–40.

36. Barnes J. Evidentials in the tuyuca verb. *International Journal of American Linguistics*, 1984, vol. 50, no. 3, pp. 255–271.

Информация об авторе

Качков Дмитрий Ильич, аспирант кафедры многопроцессорных систем и сетей факультета прикладной математики и информатики, Белорусский государственный университет.
E-mail: dmitriydikanskiy@gmail.com

Information about the author

Dzmitry I. Kachkou, Postgraduate Student of Department of Multiprocessor Systems and Networks of the Faculty of Applied Mathematics and Informatics, Belarusian State University.
E-mail: dmitriydikanskiy@gmail.com