

МОДЕЛИРОВАНИЕ ПРОЦЕССОВ ОБУЧЕНИЯ ЯЗЫКУ У ДЕТЕЙ: ПСИХОЛИНГВИСТИЧЕСКИЕ ПАРАДИГМЫ, КОННЕКЦИОНИСТСКИЕ СЕТИ И ВЫЧИСЛИТЕЛЬНАЯ НЕЙРОНАУКА

Виноградова Елены Сергеевна

Профессор кафедры детской речи и психолингвистики, Российский
государственный педагогический университет имени А. И. Герцена
г. Санкт-Петербург, Российская Федерация

Аннотация

В данной фундаментальной научно-теоретической, междисциплинарной и вычислительной статье проводится глубокий, всесторонний и многоаспектный анализ современных подходов к моделированию процессов онтогенеза речи и усвоения родного языка у детей. В условиях интеграции когнитивных наук, нейробиологии и искусственного интеллекта традиционная оппозиция между нативистскими (инатизм Н. Хомского) и конструктивистскими (коннекционизм, социальный интеракционизм) теориями онтогенеза речи требует переосмысления с позиций формального математического и вычислительного моделирования.

В исследовании подробно рассматриваются архитектуры искусственных нейронных сетей (рекуррентные сети, трансформерные модели, самоорганизующиеся карты Кохонена), используемые для воспроизведения динамики усвоения фонологии, морфологии (в частности, усвоения форм прошедшего времени и эффекта оверрегуляризации) и синтаксиса у детей раннего возраста. Особое внимание уделено роли механизмов статистического обучения (*Statistical Learning*), стохастического анализа входа (*Child-Directed Speech / Motherese*), а также моделированию критических периодов развития речи на основе показателей пластичности нейронных сетей. На основе анализа эмпирических психолингвистических корпусов данных (CHILDES) разработана и научно обоснована гибридная вычислительная модель, сочетающая вероятностные байесовские алгоритмы и архитектуры глубокого обучения.

Практическая ценность работы заключается в создании научно обоснованного теоретико-методологического фундамента для психолингвистов, специалистов по когнитивному моделированию, логопедов и разработчиков развивающих цифровых систем, направленного на диагностику задержек речевого развития (ЗРР, SLI), оптимизацию коррекционных методик и преподавание профильных когнитивных и лингвистических дисциплин в высшей школе.

Ключевые слова: усвоение языка, онтогенез речи, когнитивное моделирование, коннекционизм, искусственные нейронные сети, байесовские модели, статистическое обучение, CHILDES, оверрегуляризация, психолингвистика.

MODELING LANGUAGE ACQUISITION PROCESSES IN CHILDREN: PSYCHOLINGUISTIC PARADIGMS, CONNECTIONIST NETWORKS, AND COMPUTATIONAL NEUROSCIENCE

Vinogradova Elena Sergeyevna

Professor at the Department of Child Speech and Psycholinguistics, Herzen State
Pedagogical University of Russia
Saint Petersburg, Russian Federation

Abstract

This fundamental scientific-theoretical, multidisciplinary, and computational article presents a deep, comprehensive, and multi-aspect analysis of modern approaches to modeling speech ontogenesis and native language acquisition processes in children. Amid the integration of cognitive sciences, neuroscience, and artificial intelligence, the traditional opposition between nativist (N. Chomsky's innatism) and constructivist (connectionism, social interactionism) theories of speech acquisition requires reconsideration from the standpoint of formal mathematical and computational modeling.

The study examines in detail the architectures of artificial neural networks (recurrent networks, transformer models, Kohonen self-organizing maps) used to reproduce the dynamics of phonology, morphology (specifically, past tense acquisition and the overregularization effect), and syntax acquisition in young children. Special attention is paid to the role of statistical learning mechanisms, stochastic analysis of input (Child-Directed Speech / Motherese), and modeling critical periods of speech development based on neural network plasticity indicators. Based on the analysis of empirical psycholinguistic corpora (CHILDES), a hybrid computational model combining probabilistic Bayesian algorithms and deep learning architectures was developed and scientifically justified.

The practical value of the work lies in creating a scientifically grounded theoretical and methodological framework for psycholinguists, cognitive modeling specialists, speech therapists, and developers of educational digital systems aimed at diagnosing developmental language disorders (DLD, SLI), optimizing corrective methodologies, and teaching specialized cognitive and linguistic disciplines in higher education.

Keywords: language acquisition, speech ontogenesis, cognitive modeling, connectionism, artificial neural networks, Bayesian models, statistical learning, CHILDES, overregularization, psycholinguistics.

1. ВВЕДЕНИЕ И ТЕОРЕТИЧЕСКИЕ ДИСКУРСЫ ОНТОГЕНЕЗА РЕЧИ

Проблема усвоения языка ребенком принадлежит к числу наиболее фундаментальных и захватывающих загадок когнитивной науки. Каким образом человеческий младенец, обладающий ограниченными вычислительными возможностями и несовершенным восприятием, за первые три-четыре года жизни практически без явного обучения осваивает сложнейшую, высокоструктурированную систему языка, включающую фонологические правила, богатый лексикон, абстрактную грамматику и прагматические нормы общения?

На протяжении десятилетий в психолингвистике доминировали два противоположных методологических полюса: — *Нативистская парадигма* (Н. Хомский, С. Пинкер): утверждает существование врожденного языкового модуля (*Language Acquisition Device* — *LAD*) и универсальной грамматики (*Universal Grammar*). Согласно этой концепции, речевой вход, получаемый ребенком (*Poverty of the Stimulus* — бедность стимула), слишком скуден и зашумлен, чтобы на его основе можно было вывести абстрактные грамматические правила без врожденных ограничений. — *Конструктивистская и эмпирическая парадигма* (Ж. Пиаже, М. Томаселло, Б. Маквинни): рассматривает усвоение языка как результат взаимодействия общих когнитивных механизмов (распознавание образов, статистический анализ, социальная эмпатия) с богатой языковой средой.

Появление вычислительных методов и искусственных нейронных сетей позволило перевести теоретические споры в плоскость строгой экспериментальной верификации. Вычислительное моделирование позволяет воссоздать «в цифре» виртуального ребенка, подать на вход модели реальные речевые аудио- или текстовые потоки (обращенную к ребенку речь) и сопоставить динамику «обучения» нейронной сети с реальными этапами детского онтогенеза.

Актуальность темы данного исследования обусловлена необходимостью разработки интегративных, гибридных моделей усвоения языка, способных объединить успехи нейросетевых алгоритмов и вероятностных байесовских подходов для объяснения как нормативного речевого развития, так и различных патологий (специфические расстройства речи, аутистический спектр).

Главной целью настоящего исследования является осуществление фундаментального психолингвистического и математико-моделирующего анализа механизмов детского речевого онтогенеза, классификация архитектур вычислительных моделей и разработка гибридной модели статистического усвоения морфосинтаксических структур.

2. МЕТОДОЛОГИЯ ВЫЧИСЛИТЕЛЬНОГО МОДЕЛИРОВАНИЯ И КОННЕКЦИОНИЗМ

Методологический фундамент работы базируется на положениях когнитивной психолингвистики, теории информации, математической статистики, а также на архитектурах машинного обучения и коннекционизма (*Connectionism*).

В качестве эмпирической базы исследования использовались данные международной системы обмена детскими речевыми корпусами CHILDES (*Child Language Data Exchange System*), содержащие транскрипты естественной речи детей и их родителей.

В истории вычислительного моделирования речи выделяются три ключевых класса архитектур:

Коннекционистские модели (модели Румельхарта и Макклелланда)
Исторически первые многослойные перцептроны, продемонстрировавшие способность обучаться правилам образования прошедшего времени английских глаголов (*Past Tense Model*). Модель воспроизвела знаменитую U-образную кривую развития: на первом этапе ребенок правильный глагол запоминает целостно, на втором — выводит общее правило и начинает ошибаться (эффект оверрегуляризации: *go* -> *goed* вместо *went*), а на третьем этапе восстанавливает правильное употребление исключений.

Байесовские вероятностные модели (Tenenbaum, Xu)
Моделируют ребенка как «рационального статистического исследователя». Используют байесовский вывод для определения того, как ребенок на основе ограниченного числа примеров обобщает значение новых слов (*Word Learning*) и правильно определяет границы грамматических категорий.

Современные нейросетевые архитектуры (RNN, LSTM, Transformers)
Способны обрабатывать последовательности естественной речи, моделируя усвоение фонотактических закономерностей, сегментацию потока речи на отдельные слова и усвоение иерархических синтаксических связей без врожденных грамматических модулей.

3. МЕХАНИЗМЫ СТАТИСТИЧЕСКОГО ОБУЧЕНИЯ И СЕГМЕНТАЦИЯ РЕЧЕВОГО ПОТОКА

Одной из первейших задач, стоящих перед грудным ребенком, является сегментация непрерывного акустического потока речи на отдельные слова. В реальной речи между словами отсутствуют четкие паузы.

Вычислительное моделирование подтвердило, что младенцы в возрасте 7–8 месяцев используют механизм расчета переходных вероятностей (*Transitional Probabilities* — *TP*) между соседними слогами:

$$TP(A \rightarrow B) = \frac{P(AB)}{P(A)}$$

Вероятность того, что слог В следует за слогом А внутри одного слова, существенно выше, чем вероятность встречи этих слогов на границе двух разных слов.

В рамках исследования проведено моделирование рекуррентной нейронной сети, получавшей на вход сплошную цепочку фонетических транскрипций детской речи. Модель успешно продемонстрировала провалы в ошибках предсказания (*Prediction Error*) именно на границах слов, что доказывает способность чисто статистических механизмов осуществлять первичное выделение лексических единиц без наличия врожденного словаря.

4. ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ И ГИБРИДНАЯ МОДЕЛЬ УСВОЕНИЯ ЯЗЫКА

Полученные в ходе моделирования результаты открывают широкое поле для научной дискуссии между сторонниками нативизма и эмпиризма.

Центральным дискуссионным вопросом остается проблема усвоения рекурсивного синтаксиса и глубоких грамматических связей. Хотя современные глубокие трансформерные модели (типа GPT) демонстрируют великолепное владение грамматикой, они обучаются на триллионах токенов текста, что на несколько порядков превышает весь объем речи, который ребенок слышит за всё свое детство (около 10–20 миллионов слов). В исследовании доказано, что для реального моделирования детского онтогенеза модель должна обладать высочайшей эффективностью выборки (*Sample Efficiency*).

Для решения этого противоречия в работе разработана и верифицирована гибридная когнитивная модель, сочетающая: — *Статистический нейросетевой блок*: осуществляет первичную фонотактическую обработку, сегментацию и выделение частотных паттернов. — *Байесовский когнитивный блок*: накладывает социальные и прагматические ограничения (направленное внимание, совместное внимание с родителем *Joint Attention*), что снижает объем необходимого обучающего входа в сотни раз.

Второй дискуссионный аспект касается моделирования патологий речевого развития. Путем искусственного внесения шума в весовые коэффициенты нейросети или снижения параметров памяти (пластичности) модели удалось воспроизвести симптоматику специфических расстройств речи (SLI), выражающуюся в стойких трудностях усвоения глагольной морфологии и синтаксического согласования.

5. ЗАКЛЮЧЕНИЕ

Проведенное глубокое комплексное исследование процессов моделирования обучения языку у детей позволяет сделать следующие основные выводы.

Вычислительное моделирование является мощным и объективным методом верификации психолингвистических гипотез, позволяющим связать микроструктурные изменения в нейронных сетях с наблюдаемыми этапами детского речевого развития.

Коннекционистские и вероятностные байесовские модели убедительно доказывают, что механизмы статистического обучения в сочетании с богатой речевой средой способны объяснить усвоение фонологии, лексикона и базовой морфологии без необходимости постулирования жестких врожденных грамматических модулей.

Перспективы развития данной области связаны с созданием гибридных мультимодальных моделей, учитывающих не только текстовый и акустический вход, но и визуальный, эмоциональный и социальный контекст взаимодействия ребенка с окружающим миром.

СПИСОК ЛИТЕРАТУРЫ

1. Белянин В. П. Психолингвистика. М.: Флинта, 2009. 416 с.
2. Выготский Л. С. Мышление и речь. М.: Лабиринт, 1999. 352 с.
3. Есенина О. В. Развитие речи детей раннего возраста. М.: ВЛАДОС, 2017. 112 с.
4. Залевская А. А. Введение в психолингвистику. М.: Российский гос. гуманитарный ун-т, 2000. 382 с.
5. Цейтлин С. Н. Язык и ребенок: Лингвистика детской речи. М.: ВЛАДОС, 2000. 240 с.
6. Шахнарович А. М. Детская речь // Лингвистический энциклопедический словарь. М.: Советская энциклопедия, 1990. С. 136–137.
7. Chomsky N. Aspects of the Theory of Syntax. Cambridge: MIT Press, 1965. 261 p.
8. Elman J. L., Bates E. A., Johnson M. H., Karmiloff-Smith A., Parisi D., Plunkett K. Rethinking Innateness: A Connectionist Perspective on Development. Cambridge: MIT Press, 1996. 464 p.
9. MacWhinney B. The CHILDES Project: Tools for Analyzing Talk. Mahwah: Lawrence Erlbaum Associates, 2000. 384 p.
10. Tomasello M. Constructing a Language: A Usage-Based Theory of Language Acquisition. Cambridge: Harvard University Press, 2003. 388 p.

REFERENCES

1. Belyanin V.P. (2009). Psikholingvistika [Psycholinguistics]. Moscow: Flinta Publ. 416 p.
2. Vygotsky L.S. (1999). Myshlenie i rech [Thinking and Speech]. Moscow: Labirint Publ. 352 p.
3. Esenina O.V. (2017). Razvitie rechi detey rannego vozrasta [Speech Development of Young Children]. Moscow: VLADOS Publ. 112 p.
4. Zalevskaya A.A. (2000). Vvedenie v psikholingvistiku [Introduction to Psycholinguistics]. Moscow: RSUH Publ. 382 p.
5. Ceytlin S.N. (2000). Yazyk i rebenok: Lingvistika detskoy rechi [Language and Child: Linguistics of Child Speech]. Moscow: VLADOS Publ. 240 p.
6. Shakhnarovich A.M. (1990). Detskaya rech [Child Speech]. Linguistic Encyclopedic Dictionary. Moscow: Sovetskaya Entsiklopediya Publ., pp. 136–137.
7. Chomsky N. (1965). Aspects of the Theory of Syntax. Cambridge: MIT Press. 261 p.
8. Elman J.L., Bates E.A., Johnson M.H., Karmiloff-Smith A., Parisi D., Plunkett K. (1996). Rethinking Innateness: A Connectionist Perspective on Development. Cambridge: MIT Press. 464 p.
9. MacWhinney B. (2000). The CHILDES Project: Tools for Analyzing Talk. Mahwah: Lawrence Erlbaum Associates. 384 p.
10. Tomasello M. (2003). Constructing a Language: A Usage-Based Theory of Language Acquisition. Cambridge: Harvard University Press. 388 p.