Russian Journal of Linguistics
ISSN 2687-0088 (print), ISSN 2686-8024 (online)

2022 Vol. 26 No. 2 317–341
http://journals.rudn.ru/linguistics

https://doi.org/10.22363/2687-0088-30171

Research article

Natural language processing and discourse complexity studies

Marina SOLNYSHKINA[1] ✉, Danielle MCNAMARA[2] and Radif ZAMALETDINOV[1]

[1]*Kazan Federal University, Kazan, Russia*
[2]*Arizona State University, Tempe, USA*
✉mesoln@yandex.ru

Abstract
The study presents an overview of discursive complexology, an integral paradigm of linguistics, cognitive studies and computer linguistics aimed at defining discourse complexity. The article comprises three main parts, which successively outline views on the category of linguistic complexity, history of discursive complexology and modern methods of text complexity assessment. Distinguishing the concepts of linguistic complexity, text and discourse complexity, we recognize an absolute nature of text complexity assessment and relative nature of discourse complexity, determined by linguistic and cognitive abilities of a recipient. Founded in the 19th century, text complexity theory is still focused on defining and validating complexity predictors and criteria for text perception difficulty. We briefly characterize the five previous stages of discursive complexology: formative, classical, period of closed tests, constructive-cognitive and period of natural language processing. We also present the theoretical foundations of Coh-Metrix, an automatic analyzer, based on a five-level cognitive model of perception. Computing not only lexical and syntactic parameters, but also text level parameters, situational models and rhetorical structures, Coh-Metrix provides a high level of accuracy of discourse complexity assessment. We also show the benefits of natural language processing models and a wide range of application areas of text profilers and digital platforms such as LEXILE and ReaderBench. We view parametrization and development of complexity matrix of texts of various genres as the nearest prospect for the development of discursive complexology which may enable a higher accuracy of inter- and intra-linguistic contrastive studies, as well as automating selection and modification of texts for various pragmatic purposes.

Keywords: *text complexity, discourse, cognitive model, automatic analyzer, natural language processing*

For citation:
Solnyshkina, Marina, Danielle McNamara & Radif Zamaletdinov. 2022. Natural language processing and discourse complexity studies. *Russian Journal of Linguistics* 26 (2). 317–341. https://doi.org/10.22363/2687-0088-30171

Научная статья

Обработка естественного языка и изучение сложности дискурса

М.И. СОЛНЫШКИНА[1] ✉, Д. МАКНАМАРА[2], Р.Р. ЗАМАЛЕТДИНОВ[1]

[1]Казанский (Приволжский) федеральный университет, Казань, Россия
[2]Университет штата Аризона, Темпе, США
✉mesoln@yandex.ru

Аннотация
В исследовании представлен обзор формирования и развития дискурсивной комплексологии – интегрального научного направления, объединившего лингвистов, когнитологов и программистов, занимающихся проблемами сложности дискурса. Статья включает три основных части, в которых последовательно изложены взгляды на категорию сложности, история дискурсивной комплексологии и современные методы оценки сложности текста. Разграничивая понятия сложности языка, текста и дискурса, мы признаем абсолютный характер оценки сложности текста и относительный, зависимый от языковой личности реципиента характер сложности дискурса. Проблематика теории сложности текста, основы которой были заложены в XIX в., сфокусирована на поиске и валидации предикторов сложности и критериев трудности восприятия текста. Мы кратко характеризуем пять предыдущих этапов развития дискурсивной комплексологии: формирующего, классического, периода закрытых тестов, конструктивно-когнитивного и периода обработки естественно языка, а также подробно описываем современное состояние науки в данной области. Мы представляем теоретическую базу автоматического анализатора Coh-Metrix – пятиуровневую когнитивную модель восприятия, позволившую обеспечить высокий уровень точности оценки сложности и включить в список предикторов сложности текста не только лексические и синтаксические параметры, но и параметры текстового уровня, ситуационной модели и риторических структур. На примере нескольких инструментов (LEXILE, ReaderBench и др.) мы показываем области применения данных инструментов, включающие образование, социальную сферу, бизнес и др. Ближайшая перспектива развития дискурсивной комплексологии состоит в параметризации и создании типологии сложности текстов различных жанров для обеспечения более высокой точности меж- и внутриязыкового сопоставления, а также для автоматизации подбора текстов в различных лингвопрагматических условиях.

Ключевые слова: *сложность текста, дискурсивная комплексология, когнитивная модель, автоматический анализатор, обработка естественного языка*

Для цитирования:
Солнышкина М.И., Макнамара Д., Замалетдинов Р.Р. Обработка естественного языка и изучение сложности дискурса. *Russian Journal of Linguistics.* 2022. Т. 26. № 2. С. 317–341. https://doi.org/10.22363/2687-0088-30171

1. Введение

Более семи десятилетий исследователи в области компьютерной лингвистики искали решения, которые позволили бы компьютерным системам осуществлять обработку естественного языка. Эта работа велась как с целью решения сугубо лингвистических задач, включая, например, разработку теории общения на естественном языке (Hendrix 1980), так и для создания компьютерных систем, осуществляющих общение с пользователями на естественном языке. И хотя задача свободного понимания машиной естественного языка по-прежнему не решена, значительный прогресс достигнут в создании автоматизированных систем обработки естественного языка (см. Coh-Metrix, Lextutor, ReaderBench, RuLinva, TextInspector, Текстометр и др.), широко востребованных в системах образования и здравоохранения, социальных службах, бизнесе и праве. Потребность в такого рода инструментах заставляет исследователей искать новые алгоритмы и инструменты для решения задач, стоящих перед обществом. А сложность исследовательской задачи – автоматизации лингвистического анализа – обусловлена сложностью самой языковой системы, процесса восприятия вербальной информации в целом и в текстовом формате в частности, а также многообразием текстовых типов и жанров. В известной работе «Науки об искусственном» (1996) Г. Саймон, определяя основную задачу науки в целом, фактически обозначил направление развития проблематики сложности текста: «сделать удивительное тривиальным, показать, что правильно рассматриваемая сложность является лишь маской простоты (англ. complexity, correctly viewed, is only a mask for simplicity), найти закономерность, спрятанную в кажущемся хаосе»[1] (Simon 1996: 1).

В представленной работе мы предлагаем свой взгляд на категорию сложности текста, а также кратко описываем историю и достижения лингвистической комплексологии, реализованные в автоматических анализаторах различного уровня и класса.

2. Проблема сложности: сложность языка, текста и дискурса

Сложность признана ключевой характеристикой бытия, и стремление понять сложные системы объединяет ученых различных направлений. Новым в современной исследовательской парадигме является не изучение отдельных сложных систем, но описание самого феномена сложности, а также способов и инструментов оценки сложности. Термин «сложность» определяется в науке при помощи широкого спектра понятий: «многообразность»,

[1] Здесь и далее перевод с английского выполнен авторами статьи.

«множественность», «иерархия», «неоднородность», «неаддитивность» или «эмерджентность», «холизм», «гештальт», «информация», «общие системы», «генетические алгоритмы» и др. Например, Н. Решер определяет сущность сложности через «(1) количество и многообразие дискретных компонентов, из которых состоит объект, а также (2) количество связей между составляющими компонентами» (Rescher 1998: 1). Г. Саймон пишет о «неаддитивности» или «эмерджентности» сложных систем, их несводимости к сложности элементов: «Под сложной системой мы понимаем систему, состоящую из большого числа частей и взаимодействующую между собой непростым образом. В таких системах целое больше, чем сумма частей, в том смысле, что по заданным свойствам частей и их взаимодействиям нельзя правильным образом выявить свойства всей системы» (Саймон 2004: 104–105). Ф. Андерсон указывает на иерархическую структуру сложных объектов (Anderson 1972), а Г. Саймон подчеркивает, что «…сложность часто проявляется в форме иерархии и что иерархические системы имеют некоторые общие свойства, не зависящие от их конкретного содержания. <…> иерархия – это одна из центральных структурных схем, которую использует архитектор сложности» (Simon 1996: 184).

Что касается лингвистической сложности, то в современной научной парадигме сформировано три взаимосвязанных понятия: сложность языка (англ. language complexity), сложность текста (англ. text complexity) и сложность дискурса (англ. discourse complexity), а термин «лингвистическая сложность» (англ. linguistic complexity) используется для обозначения каждого из вышеуказанных понятий (см. Kortmann & Szmrecsanyi 2012). В некоторых случаях и термин «сложность языка» используется как синоним термина «сложность текста» (Sun 2020). Рассмотрим каждое из этих понятий. Наибольшую известность получили относительный и абсолютный подходы к понятию сложности, в рамках которых относительная и абсолютная сложность трактуются как свойство функций языка (т.е. элементов, паттернов, конструкций, правил), их (под-)систем или способов использования этих функций. Однако исследователи расходятся во мнениях относительно того, можно ли говорить об абсолютной сложности языка, т.е. существует ли некая независимая мера, с помощью которой оценивается сложность любого языка, либо сложность языка следует оценивать только в сравнении с другим языком. Можно ли оценивать сложность/простоту языка в целом (холистическая сложность) или, поскольку различные параметры сложности объективируются в разных структурах, оценка возможна только для отдельных уровней (уровневая сложность)? Весьма болезненным является вопрос о «количественной», т.е. структурной сложности языка, когда элементам или категориям в одном языке присваивается более высокая степень сложности, чем их аналогам в других языках. Поскольку при оценке абсолютной количественной сложности качественные аспекты сложности, такие как внутренняя сложность отдельных категорий, функций, правил, не учитываются, развитие

мысли о более высокой относительной сложности одного языка по сравнению с другим без учета степени «когнитивной сложности» языковых фактов может привести к выводу о том, что структурная простота, например изолирующих языков, означает примитивность развития их носителей (Steger & Schneider 2012: 159).

При оценке *сложности текста* разрабатываются шкалы уровней сложности текстов, предназначенных для различных категорий реципиентов, а присвоению тексту определенного уровня сложности предшествует создание «инвентарного» списка элементов, составляющих сложность каждого уровня. В этом случае текст оценивается как более или менее сложный на основании присутствия/отсутствия в тексте элементов, объективирующих «сложности» соответствующей шкалы. Несмотря на относительный характер *сложности текста* и самой процедуры, речь в данном случае идет об объективации «списочного подхода», т.е. абсолютной сложности. А относительной сложностью принято называть индивидуальную трудность текста для отдельного читателя. К сожалению, в ряде случаев индивидуальную трудность ошибочно называют субъективной сложностью (Bulté & Housen 2012: 31–33), хотя для читателя она весьма объективна.

И хотя достижения когнитологов и специалистов в области искусственного интеллекта позволили реализовать сложные модели верификации предикторов сложности и ввести в список обязательных семантические и дискурсивные параметры, термины «сложность текста» и «сложность дискурса» в современных исследованиях до сих пор используются как синонимы, обозначающие перечень параметров текста, детерминирующих трудность его восприятия читателями. Очевидно, что ускользающая от исследователя сущность лингвистической сложности текста и дискурса детерминирована в первую очередь неоднозначностью самих объектов, а также многообразием подходов к данным феноменам. Обратимся к истории разработки проблематики лингвистической сложности.

3. История автоматизации оценки сложности текста и дискурса

Приступая к краткому изложению истории автоматизации оценки сложности, следует отметить, что почти за 30 лет до появления первых формул читабельности русский ученый Н.А. Рубакин опубликовал свой знаменитый труд «Заметки по литературе для народа» (Рубакин 1890), в котором убедительно доказал, что сложность текста зависит от «знакомости» слов и длины предложения. А в 1893 г. вышла в свет работа профессора Л.А. Шермана «Analytics of Literature» («Аналитика литературы») (Sherman 1893), в которой ученый сформулировал два и по сей день никем не опровергнутых предиктора сложности текста: длина предложения и степень абстрактности. Еще одним важным событием, предопределившим появление первой формулы читабельности, следует признать идею использования частот слов в качестве параметра сложности. Именно реализация этой идеи в частотном списке слов Е. Торндайка (1921) обеспечила основу первой формулы читабельности.

Принцип, лежащий в основе частотных списков, заключается в том, что именно частота слова влияет на его узнавание в тексте, а значит, и на легкость/трудность восприятия информации в тексте.

Что касается автоматизации оценки сложности текста, то начало данного процесса авторы обзоров различных периодов развития комплексологии (см. Collins-Thompson 2015) традиционно связывают с именами Б. Лайвли и С. Прессли, разработавшими в 1923 г. первую формулу сложности текста на английском языке (Lively & Pressey 1923). Важным вкладом в теорию текстовой комплексологии того времени стали разработки М. Фогеля и К. Вошборна (Vogel & Washburne 1928), активно использовавшиеся практически без изменений до начала XXI в. Отправной точкой их исследования стало установление корреляции предикторов сложности текста с выбранными ими критериями: тестом на понимание, продолжительностью чтения текста, экспертной оценкой. В список предикторов сложности вошли лексические и синтаксические параметры, доли различных частей речи в тексте, а также информация о структуре абзаца и книги. Формула сложности М. Фогеля и К. Вошборна для английского языка имела следующий вид: $X_1 = 17{,}43 + 0{,}085X_2 + 0{,}101X_3 + 0{,}604X_4 – 0{,}411X_5$, где X_1 – сложность текста; X_2 – количество разных слов в выборке из 1000 слов; X_3 – количество предлогов в тексте; X_4 – количество слов, отсутствующих в списке Торндайка (1921); X_5 – количество простых предложений в выборке из 75 предложений. Первый этап развития дискурсивной комплексологии – формирующий – продолжался вплоть до 1935 г. и ознаменовался четырьмя важными достижениями: адаптацией способов оценки сложности, разработанных на детской учебной литературе, для других категорий читателей; включением в список предикторов «плотности идей» в тексте (Ojemann 1934); расширением списка критериев сложности текста и введением в него скорости чтения (McClusky 1934), а также включением 44 переменных в расчеты сложности текста (Gray & Leary 1935). Показательно, что уже в первых работах по сложности исследователи обращаются к широкому спектру параметров текста. Более того, именно в этот период, в 1925 г., вышел в свет первый сборник стандартизированных тестов по чтению У. МакКолла и Л. Краббс (McCall & Crabbs 1925), включавший 376 текстов и созданных на их основе тестов с множественным выбором. В течение длительного времени их валидность, установленная в ходе экспериментальных исследований с 2000 школьников Нью-Йорка, не подвергалась сомнению, а ранжированные тексты и тесты использовались в качестве критериев степени сложности. Надежность этого сборника, а вместе с тем и валидность основанных на нем формул впервые были подвергнуты жесткой критике в статье К.С. Стивенс, вышедшей в 1980 г. (Stevens 1980). Автор заявила, что инструмент плохо стандартизирован и никогда не предназначался для тех исследовательских целей, в которых он использовался.

Второй период развития комплексологии – *классический* (Klare 1963). Характерной чертой этого периода стало стремление к простоте и эффективности, нашедшее выражение в попытке разработать универсальные формулы

читабельности текстов, основанные на двух-трех предикторах. Попытки сократить количество переменных в формулах были обусловлены еще и отсутствием автоматизированных процедур расчетов переменных, что делало процедуру весьма утомительной. Именно в этот период появились простые в использовании и ставшие впоследствии классическими формулы: формулы Р. Флеша (Flesch 1948), предназначенной для взрослых, и формулы Е. Дейла и Дж. Чолл (Dale & Chall 1948), предназначенной для школьников. Обе формулы имели в своем составе два предиктора: синтаксический (среднее количество слов в предложении) и лексический (среднее количество слогов на 100 слов в формуле Р. Флеша и количество слов, не входящих в список Е. Торндайка, для формулы Е. Дейла и Дж. Чолл). Эти два исследования вдохновили множество последователей, сохранивших оригинальную методологию: использование в качестве предикторов не более двух переменных (лексических или синтаксических), а в качестве критериев – стандартизированные тексты и тесты У. МакКолла и Л. Краббс (см. McCall & Crabbs 1925).

Точкой отсчета следующего периода – периода так называемых *закрытых тестов*[2], пришедшего на смену классическому, стал выход в 1953 г. статьи У. Тейлора «Закрытый тест: новый инструмент для измерения читабельности» (Taylor 1953). В 1965 г. была опубликована книга Е. Коулмэна, в которой автор предлагал использовать закрытые тесты в качестве нового критерия сложности, утверждая, что количество правильно заполненных читателем пробелов является более достоверным критерием восприятия текста, а следовательно, и его сложности, чем ответы на вопросы по тексту (Coleman 1965). Формулы, рассчитанные на основе этого критерия, показали более эффективные результаты. Например, если коэффициенты множественной корреляции (R), полученные по формулам Р. Флеша или Е. Дейла и Дж. Чолл, не превышали 0,7, то результаты достоверности исследований Е. Коулмэна (1965) находились в пределах от 0,86 до 0,9, а Дж. Бормута (Bormuth 1969) – составляли более 0,92.

Еще одной особенностью этого периода стало то, что формулы на данном этапе «революции закрытого теста» (англ. the revolution of the cloze) включали больше предикторов, чем классические формулы. Этому способствовало появление и активное использование компьютеров, которые позволили исследователям автоматизировать подсчет некоторых переменных и обучение статистических моделей. В 1961 г. Э. Смит опубликовал первую формулу, расчеты которой были полностью автоматизированы (Smith & Quackenbush 1960), – формулу Деверо (Devereux). В качестве параметров Э. Смит использовал количество символов (знаков) в слове, аргументируя свой выбор тем, что это проще и быстрее, чем считать слоги или выбирать слова из списка (Smith & Quackenbush 1960: 333). Уже в 1963 г. У. Дэнилсон и С. Брайан адаптировали формулу Э. Смита для расчета на компьютере UNIVAC 1105 (Danielson & Bryan 1963). Одним из выдающихся исследователей того

[2] Cloze test – тест на заполнение пропусков в связном тексте.

периода был Дж. Бормут (Bormuth 1969). Именно Дж. Бормут впервые обратился к ряду методологических вопросов текстовой комплексологии: он первым показал, что связь между предикторами и критериями сложности не всегда является линейной. Он первым использовал дерево синтаксического анализа в качестве предиктора сложности и доказал, что оно менее эффективно, чем количество слов в предложении. Ему принадлежит и первенство расчета сложности на трех уровнях: уровне текста, предложения и слова. Он подвергает критике исследования, в которых рассчитывается коэффициент корреляции не из тестового, а обучающего набора данных.

В конце XX в. интерес к проблематике сложности текста наблюдается в отечественной науке изучается сложность текстов различных стилей и жанров, выходят ряд публикаций, появляются формулы читабельности для русского языка. В качестве переменных используются исключительно лексические и синтаксические предикторы: доля многозначных слов, абстрактных слов, трехсложных слов, средняя длина предложений и некоторые др. (см. обзор Солнышкина, Кисельников 2015). К сожалению, возможность автоматизации большого количества параметров привела к тому, что исследователи в ряде случаев стали выбирать параметры текста, расчет которых не представлял особых сложностей. В конечном итоге это способствовало формированию более поверхностного взгляда на сложность текста. Именно поэтому, а также благодаря достижениям когнитологов в области изучения процессов восприятия информации новые формулы читабельности подвергались все более резкой критике (Kintsch & Vipond 1979). Основные замечания когнитологов были нацелены на неспособность классических формул учитывать связность, когерентность, композицию и другие характеристики текста, а также интерактивность самого процесса чтения. Кроме того, дискредитации формул способствовала и повсеместная практика их применения на текстах таких типов, критерии сложности которых не были изучены. Практика такого рода приводила к недостоверности расчетов. В это же время появляются работы, авторы которых заявляют, что «смыслы несут люди, но не тексты» (Spivey 1987: 171), а процесс чтения и оценка сложности текста требуют учета индивидуальных способностей читателя – памяти, мотивации и др. На смену периоду «закрытых тестов» пришел новый – *структурно-когнитивный* или *когнитивный*. Появившись как следствие достижений в области лингвистики и когнитивных наук, он стал прорывом в лингвистической комплексологии: критика формул читабельности сопровождалась появлением исследований с новыми предикторами, измеряющими сложные психолингвистические характеристики, такие как информационная нагрузка текста (англ. inferential load) (Kemper 1983), концептуальная плотность текста (англ. conceptual density) (Kintsch & Vipond 1979) или композиция и тип текста (Meyer 1982). Следует, однако, признать, что, несмотря на гораздо более сложные предикторы, сочетание структурно-когнитивистских характеристик с классическими переменными привело к минимальному улучшению достоверности

(R = 0,78 vs 0,76) (см. Kemper 1983). Привлечение психолингвистических и даже нейролингвистических данных в комплексологию текста позволило поднять исследования на новый уровень, а «погружение» текста в коммуникативную ситуацию повлекло за собой включение в осуществляемые исследования сложности двух новых переменных: ситуативного контекста и языковой личности читателя. Текстовая комплексология стала дискурсивной комплексологией.

В 1990–2000-е гг. продолжают разрабатываться формулы, имеющие в своей основе традиционный подход: например, формула Дж. Чолл и Е. Дейла 1995 г. по сути является всего лишь обновленным вариантом старой формулы (Chall & Dale 1995). В это же время появляется серьезный интерес к данной проблематике в России и формулы сложности текстов на русском языке: формула И.В. Оборневой для художественных текстов и формула SIS (Solovyev-Ivanov-Solnyshkina) для учебных текстов (см. Solovyev et al. 2018, Solnyshkina et al. 2020). Именно в эти годы были заложены успехи современного этапа развития дискурсивной комплексологии – этапа *искусственного интеллекта*. Фокус парадигмы того времени был нацелен не только на поиск новых методов оценки сложности объекта, но и на создание аналитических инструментов, способных оптимально быстро дать характеристику текста и осуществить анализ его сложности. Формулируя требования к инструментам оценки сложности в 1993 г., М. Рабин подчеркивает значимость их способности рассчитывать три параметра: (1) длину оцениваемых последовательностей внутри системы, определяющую время их оценки; (2) «глубину вычисления», т.е. число уровней параллельных шагов, на которые последовательность может быть разложена; (3) объем памяти, требуемый для расчетов (Рабин 1993: 382). Именно в эти годы была создана система SATOdication, позволявшая осуществить автоматическую оценку 120 лингвистических переменных с весьма высоким коэффициентом корреляции расчетов, достигавшим отметки R = 0,86 (Daoust et al. 1996). Вскоре после этого была предложена новая метрика лексической связности текста, основанная на латентном семантическом анализе (LSA). Латентный семантический анализ может проводиться на семантически размеченных больших корпусах данных. П. Фольц и др. (Foltz et al. 1998) доказали, что предложения, расположенные в одном семантическом пространстве корпуса, имеют высокую вероятность внутренних связей. Это свойство стало использоваться для оценки глобальной когерентности текста и рассчитываться как среднее косинусное сходство всех пар смежных предложений. Для дискурсивной комплексологии это открытие оказалось весьма значимым: было доказано, что это – новый предиктор, поскольку лексическая связность коррелирует с критерием сложности текста.

А уже в 2001 г. Л. Си и Дж. Каллан определяют сложности текста как проблему классификации, расширяя таким образом спектр используемых для оценки методов, и применяют методы машинного обучения. Ученым удалось

выявить корреляции языковых моделей униграмм с типичными для разного уровня сложности контентами. Классификаторы уровней сложности при этом формируются как линейные комбинации языковой модели и поверхностных лингвистических признаков (Si & Callan 2001).

Существенный вклад в развитие теории сложности текста внесли исследования Ф. Даоста и др. (Daoust et al. 1996), П. Фольца и др. (Foltz et al. 1998), Л. Си и Дж. Каллан (Si & Callan 2001), которые и фактически поменяли научную парадигму. Исследования в новой парадигме по-прежнему предполагают наличие размеченного корпуса и выбор предикторов, анализ и комбинации которых верифицируются в статистических моделях, однако использование методов обработки естественного языка позволяет автоматизировать расчеты большого количества предложенных в рамках структурно-когнитивного подхода параметров. В первое десятилетие нашего века появляются исследования, нацеленные на валидацию не только лексических и синтаксических предикторов, но в первую очередь семантических, дискурсивных и когнитивных. Например, на основе расчетов, сделанных при помощи профайлера Coh-Metrix, исследовательская группа в составе С. Кроссли, Д. Дафти, П. МакКарти и Д. МакНамара разработала первую формулу сложности текста, сочетающую лексические и синтаксические параметры с параметрами связности (Crossley et al. 2007). Приблизительно в эти же годы исследователи делают еще одно важное открытие: синтаксические модели являются предиктором сложности текста для неносителей языка, в то время как на восприятие носителей языка сложность синтаксических моделей существенного влияния не оказывает (Schwarm & Ostendorf 2005).

Современные исследования в области дискурсивной комплексологии показывают, что эффективными моделями сложности текста являются только модели, ориентированные на конкретную категорию реципиентов, а выявление корреляций предикторов и критериев сложности предполагает учет целевой аудитории для конкретного текста. Фактически открытие этого важного закона предопределило качественно новый этап в развитии теории сложности и появление *дискурсивной комплексологии*, в которой, в отличие от комплексологии текста, критерии сложности текста выстраиваются в зависимости от категорий реципиентов. Два основных вызова времени – недостаточно качественная аннотация корпусов, искажающая данные, и необходимость уточнения и корректировки критериев сложности на различных категориях читателей, включая читателей с нарушениями речи, а также носителей и неносителей языка – решаются в новейшей истории комплексологии по-разному. Сочетание машинного обучения и методов обработки естественного языка должны обеспечивать лучшую точность предикторов, однако для современных более сложных статистических моделей нужно большее (по сравнению с предыдущими периодами) количество параметров для их обучения, следовательно, значительно возрастает потребность в больших, хорошо размеченных обучающих корпусах. Именно поэтому ученые все чаще обращаются к уже

размеченным корпусам, т.е. к корпусам с уже присвоенными уровнями сложности. Это может быть, например, корпус линейки учебников, сложность которых возрастает, например, в диапазоне от 2-го до 11-го класса (Solovyev et al. 2019, Gatiyatullina et al. 2020); корпус текстов для изучающих язык как неродной, в котором каждому тексту присвоен уровень сложности по Общеевропейской шкале (Лапошина, Лебедева 2021). Современный период комплексологии – период междисциплинарных исследований, в которых для изучения сложности привлекаются лингвисты, когнитологи, программисты и специалисты в области статистики. В настоящее время совершенно очевидно, что модели оценки сложности текста должны строиться на типологии и параметризации текстов различных регистров и жанров, типологии когнитивных моделей восприятия текста, а также на типологии реципиентов.

4. Инструменты оценки сложности текста и дискурса

В начале нового века появилась потребность в обработке больших массивов данных: компании, организации и государственные органы все чаще стали испытывать необходимость в анализе информации, полученной из нескольких текстовых источников, таких как онлайн-обзоры, электронные письма, транскрипты встреч и конференций или, например, приложений для обмена сообщениями. В настоящее время столь сложные задачи вполне осуществимы с помощью методов текстовой аналитики, а известные примеры практического применения методов обработки естественного языка включают автоматизированный машинный перевод, системы ответов на вопросы, извлечения и интерпретации информации, а также анализа настроений. Как ответ на практические запросы общества с конца прошлого века начинают появляться инструменты (системы, анализаторы, профайлеры) автоматического анализа текстов.

Появление автоматических анализаторов текстов не решило всех стоящих перед учеными проблем, но в определенной степени обострило их. Известно, что разработка и функционирование инструментов автоматического анализа естественного языка во многом зависят от имеющихся в распоряжении конструкторов баз данных, способов доступа к ним, а также средств и способов представления данных. Именно базы данных являются основой, на которой может строиться работа любой автоматизированной системы обработки естественного языка. При этом очевидно, что информация, доступная даже в нескольких базах данных, как правило, довольно ограничена (см. об этом Соловьев и др. 2022). Более того, пользователи, которым предоставляется доступ к базе данных, во многих случаях ожидают не только извлечения нужной им информации, хранящейся в базе, но и способности системы осуществлять на этих данных расчеты производной информации. При разработке эффективных систем автоматизированной обработки языка возникает также проблема создания средств представления данных. По мере того как компьютерные системы используют более сложные базы знаний, им

требуются более эффективные средства представления данных. Извлечение информации из текстов предполагает последующую ее визуализацию в виде таблиц, карт памяти, диаграмм, которые позднее интегрируются в базы данных и используются для описательной, предписывающей или прогнозной аналитики. Однако для того, чтобы система могла вести диалог с пользователем, ожидается, что она не только умеет интерпретировать вводимые данные, но и соответствующим образом реагирует на запросы пользователя, генерируя ответы, уже адаптированные к предполагаемым потребностям пользователя. Иллюстрацией решения такого рода проблемы в области обработки естественного языка является создание разработчиками Coh-Metrix «сокращенного» и более удобного в пользовании приложения TERA (Text Ease and Readability Assessor, букв. Анализатор легкости и удобочитаемости текста), предназначенного преимущественно для педагогов (ENA, April 18, 2022)[3]. TERA рассчитывает и визуализирует интегральные индексы пяти кластеров параметров: повествовательность, синтаксическая простота, конкретность слова, референциальная связность и глубокая связность, востребованных методистами и тестологами при отборе текстов для соответствующей аудитории читателей (см. Solnyshkina, Harkova & Kiselnikov 2014). Coh-Metrix же анализирует тексты по более чем 200 параметрам и предназначена преимущественно для исследовательских целей (McNamara & Graesser 2012).

Среди наиболее известных и хорошо зарекомендовавших себя инструментов для анализа текстов выделим следующие: ATOS (Advantage/TASA Open Standard), Lexile Framework (Lexile), REAP, проект университета Карнеги-Меллон, TextIspector, проект Открытого университета, Coh-Metrix, проект университетов Мемфиса и Штата Аризоны, США и Readerbench (ENA, April 18, 2022)[4], многоязычный профайлер Политехнического университета Бухареста. Для текстов на русском языке созданы две успешно функционирующие системы: Текстметр[5] (ENA, April 18, 2022), онлайн-инструмент определения уровня сложности текста РКИ, и RuLingva (ENA, April 18, 2022)[6], функционал которой поддерживает учебные тексты для носителей и неносителей русского языка.

В 2000-е гг. в Институте школьного возрождения (School Renaissance Institute) и компании Touchstone Applied Science Associates[7] была создана платформа ATOS (Advantage-TASA Open Standard), осуществляющая оценку текста по шести параметрам: количество слов в предложении, символов в слове, слогов в слове, частотность слова, процент знакомых слов и год овладения словом (ATOS Readability Formula). При разработке платформы

[3] http://129.219.222.70:8084/Coh-Metrix.aspx

[4] http://www.readerbench.com/

[5] https://textometr.ru/

[6] https://rulingva.kpfu.ru/

[7] В марте 2007 г. компания была переименована в Questar Assessment, Inc. (https://www.questarai.com/).

был создан корпус объемом 474 млн слов, 28 тыс. книг, 650 стандартизированных текстов, предназначенных для читателей нескольких уровней образования. В проекте были использованы записи чтения более 30 тыс. участников исследования. Параметр «год овладения словом» оценивался на основе ранжированного по годам обучения списка слов (Graded Vocabulary List). Список создан на основе «Словаря живых слов» (Living Word Vocabulary) (Dale & O'Rourke 1981), «Руководства по частотности слов» (Educator's Word Frequency Guide) (Zeno et al. 1995), «Списка слов для обучения (Words Worth Teaching) (Biemiller 2009), а также слов, используемых в стандартных тестах. Для удобства пользователей сложность текста шкалируется по году обучения в диапазоне от 0 до 15+. На платформе ATOS также были размещены списки книг, ранжированных по уровням сложности (Renaissance, Development of the ATOS, Special Collections. Accelerated Reader).

В современной версии анализатора Lexile, запущенного также в 2000-е гг. на платформе Lexile Platform (Lexile Framework for Reading), оценка сложности текста осуществляется при помощи двух индексов: семантического и синтаксического. Первый имеет в своей основе меру «незнакомости» слова, рассчитываемую при помощи частотности слов в корпусе, а второй – среднюю длину предложения (Lennon & Burdick 2004). Платформа активно используется в англоязычных странах для подбора учебников, тестовых материалов, а также развивающих книг для различных категорий читателей.

При создании анализатора DRP (Degrees of Reading Power, степени читательской способности), разработанного для платформы Questar (ENA, April 18, 2022) [8], была использована формула сложности Дж. Бормута с четырьмя предикторами: длина слова и предложения, количество знаков препинания и «узнаваемость» слова (Nelson et al. 2012). Сложность текста DRP ранжируется на непрерывной шкале от 0 до 100: более высокие значения маркируют более сложные тексты.

Отдельную нишу среди анализаторов текстов занимают так называемые «поисковые роботы» (англ. web crawlers), предназначенные для поиска вебтекстов определенной тематики и уровня сложности. Из наиболее известных укажем на REAP (Heilman et al. 2008) и TExtEvaluator (Sheehan et al. 2014). Профайлер REAP (REAder-specific Practice, букв. читательские практики) (ENA, April 18,2022)[9] основан на оценке сложности отдельных слов и не оценивает параметры высоких уровней, такие как семантика текста и связность. Ключевым компонентом REAP является расширенная модель поиска, способная находить документы, удовлетворяющие набору разнообразных лингвопрагматических запросов, включая тему текста, уровень образования (год обучения), тип синтаксических структур и словарный запас, соответствующий определенному уровню образования. REAP использует тематический классификатор SVM (Support Vector Machine, букв. метод опорных векторов).

[8] https://www.questarai.com

[9] https://www.lti.cs.cmu.edu/projects/language-technologies-education/reap-reader-specific-lexical-practice-improved-reading

Поиск выполняется на коллекции автоматически собранных из интернета документов, каждый из которых аннотирован и прошел процедуру метаразметки (Heilman et al. 2008).

TextEvaluator, заменивший SourceRater (ENA, April 18, 2022)[10], автоматизированный профайлер сложности текста, разработанный компанией ETS, имеет аналогичный функционал и адресован методистам, издателям учебников и разработчикам тестов для отбора текстов, соответствующих рекомендациям по сложности текста.

Преимущественное большинство инструментов автоматического анализа создано для английского языка, однако в последнее время стало появляться все больше инструментов для других языков: французского (François & Naets 2011), румынского (Dascalu 2014), итальянского (READ-IT), португальского (Antunes 2019, Marujo et al. 2009), русского (Gatiyatullina et al. 2020), голландского (readability 0.3.1, ENA, April 18, 2022)[11] языков.

Разработчики всех без исключения автоматических анализаторов признают, что обработка текстов пьес, интервью, стихов, рецептов или даже списков с нестандартной пунктуацией дает весьма противоречивые результаты. Наиболее существенным ограничением для современных систем обработки естественного языка считается буквальность интерпретации смыслов, т.е. неспособность систем осуществлять обработку переносных смыслов. Однако область применения профайлеров весьма широка и включает социальную сферу (Vergara & Lintao 2020), систему образования, медицину (Antunes 2019), юриспруденцию (Hall et al. 2006) и ряд других. Во всех этих сферах автоматические анализаторы применяются прежде всего для обеспечения так называемых «дискурсивных ограничений» (language-discourse constraints) (ENA, April 18, 2022)[12], т.е. ограничений на использование определенных языковых единиц в выбранных типах текстов.

5. Coh-Metrix: реализация подходов к оценке сложности дискурса

Достижения современной парадигмы дискурсивной комплексологии наилучшим образом реализованы в автоматическом анализаторе Coh-Metrix, предназначенном для оценки языковой сложности и тестирования когнитивных моделей восприятия. Инструмент прошел проверку временем и валидирован в ряде исследований (Hall et al. 2006, Graesser et al. 2014, Solnyshkina et al. 2014, Солнышкина, Кисельников 2015). Coh-Metrix создавалась на корпусе TASA (Touchstone Applied Science Associates), репрезентативность которого обеспечена не только его объемом (37 000 текстов), но и тем, что он «содержит примерно такое же количество и качество текстов, которые среднестатистический студент колледжа прочел за всю свою жизнь» (Jones et al.).

[10] https://texteval-pilot.ets.org/TextEvaluator/Default2.aspx; https://www.ets.org/research/policy_research_reports/publications/report/2015/junz

[11] https://pypi.org/project/readability/

[12] https://benjamins.com/catalog/pbns.172

Coh-Metrix осуществляет глубокую обработку текста на лексическом и синтаксическом уровнях, определяет степень связности текста и дискурсивные параметры текста, эксплицитность ситуационной модели. В основу разработки Coh-Metrix положена пятиуровневая когнитивная модель восприятия текста (Graesser & McNamara 2011) (см. табл. 1). Характеристика данной модели выходит за пределы данной статьи (см. Солнышкина и др. 2022), однако следует указать, что для восприятия представленной в тексте информации реципиент должен иметь способность понимать отдельные слова, извлекать информацию из долговременной памяти, использовать синтаксические знания для синтеза пропозиций и макропропозиций, применять эксплицитные и имплицитные коннекторы для установления связей между предложениями и частями текста, а также использовать фоновые знания и опыт для формирования когерентной модели текста (т.е. ситуационной модели (по Кинчу)).

Таблица 1. **Уровни восприятия дискурса /**
Table 1. **Levels of discourse (Graesser & McNamara 2011)**

Levels of discourse	Уровни восприятия дискурса
(1) Surface code Word composition (graphemes, phonemes, syllables, morphemes, lemmas, tense, aspect) Words (lexical items) Part of speech categories (noun, verb, adjective, adverb, determiner, connective) Syntactic composition (noun-phrase, verb-phrase, prepositional phrases, clause) Linguistic style and dialect	**(1) Параметры поверхностного кода** Состав слова (графемы, фонемы, слоги, морфемы, леммы, время, вид) Слова (лексические единицы) Частеречные категории (существительное, глагол, прилагательное, наречие, модификатор, связка) Синтаксис (именная группа, глагольная конструкция, предложная группа, предложение) Языковой стиль и диалект
(2) Text base Explicit propositions Referents linked to referring expressions Connectives that explicitly link clauses Constituents in the discourse focus versus linguistic presuppositions	**(2) Уровень текста** Эксплицитно выраженные пропозиции Референты и их связи с наименованиями референтов Эксплицитно выраженные коннекторы частей предложения Фокус дискурса и языковые факты
(3) Situation model Agents, objects, and abstract entities Dimensions of temporality, spatiality, causality, intentionality Inferences that bridge and elaborate ideas Given versus new information Images and mental simulations of events Mental models of the situation	**(3) Ситуационная модель** Агенты, объекты и абстрактные сущности Измерения темпоральности, пространственности, причинности, интенциональности Умозаключения, обеспечивающие связь и уточнение основных идей Данная и новая информация Образы и ментальные симуляции событий Ментальные модели ситуации
(4) Genre and rhetorical structure Discourse category (narrative, persuasive, expository, descriptive) Rhetorical composition (plot structure, claim + evidence, problem + solution, etc.)	**(4) Жанр и дискурсивные параметры текста** Категории дискурса (типы текстов) (повествовательный, аргументативный, информативный, описательный)

Levels of discourse	Уровни восприятия дискурса
Epistemological status of propositions and clauses (claim, evidence, warrant, hypothesis) Speech act categories (assertion, question, command, promise, indirect request, greeting, expressive evaluation) Theme, moral, or point of discourse	Структура текста (структура сюжета, заявление + доказательство, проблема + решение и т.д.) Эпистемологический статус суждений и предложений (утверждение, доказательство, предписание, гипотеза) Категории речевого акта (утверждение, вопрос, команда, обещание, косвенная просьба, приветствие, экспрессивная оценка) Тема, мораль или идея
(5) Pragmatic communication Goals of speaker/ writer and listener/reader Attitudes (humor, sarcasm, eulogy, deprecation) Requests for clarification and backchannel feedback (spoken only)	**(5) Прагматический уровень коммуникации** Цели говорящего/писателя и слушателя/читателя Отношение (юмор, сарказм, восхваление, осуждение) Запросы разъяснений и обратной связи по обратному каналу (только в устной форме)

Рассмотрим, каким образом и при помощи каких параметров Coh-Metrix осуществляет анализ текста на английском языке.

Параметры поверхностного кода. Лексические и семантические метрики оцениваются в Coh-Metrix на основе баз данных и корпусов, включая одну из наиболее полных – MRC (ENA, April18, 2022)[13]. Coh-Metrix рассчитывает многозначность слов, возраст освоения, образность, абстрактность, долю знаменательных слов в тексте и др. (см. McNamara et al. 2014: 247–251). Частота слов рассчитывается на основе данных CELEX (ENA, April 18, 2022)[14], неоднозначность и «гиперонимический уровень»[15] – на основе WordNet (ENA, April 18, 2022)[16].

Синтаксическая сложность текста рассчитывается при помощи синтаксического анализатора Е. Чарняк (2000) (Charniak 2000: 132–139) на основе количества именных и глагольных групп, предложных словосочетаний, встроенных конструкций, модификаторов именных групп, слов перед сказуемым в главном предложении и др. Полный перечень индексов синтаксической сложности представлен в (McNamara et al. 2014: 247–251).

Уровень текста содержит эксплицитно выраженные пропозиции, референциальные связи и подтексты (англ. inferences), необходимые для понимания (van Dijk & Kintsch 1983). В качестве иллюстрации А. Грейсер

[13] https://websites.psychology.uwa.edu.au/school/mrcdatabase/mrc2.html

[14] https://www.ldc.upenn.edu/language-resources/data/obtaining

[15] «Гиперонимический уровень» есть способность слова вступать в гиперо-гипонимические отношения, коррелирующая со степенью абстрактности слова. Например, существительное *стол* имеет семь гиперонимических уровней: место – мебель – обстановка – инструментальность – артефакт – объект – сущность и обладает высокой степенью конкретности. Абстрактные слова, например, humanity, имеют более низкий «гиперонимический уровень»: humaneness (человечность) – quality (качество) – attribute (признак) – abstraction, abstract entity (абстракция, абстрактная сущность) – entity (сущность) (http://wordnetweb.princeton.edu/perl/webwn).

[16] https://wordnet.princeton.edu/

и Д. Макнамара (Graesser & McNamara 2011) предлагают следующее предложение и его пропозициональный анализ:

> When the board met on Friday, they discovered they were bankrupt. They needed to take some action, so they fired the president. *Букв. Когда совет собрался в пятницу, они поняли, что обанкротились. Им нужно было принимать меры, поэтому они уволили президента.*

Первое предложение содержит следующие пропозиции:

PROP 1: meet (board, TIME = Friday)	ПРОП-Я 1: встречаться (совет, ВРЕМЯ = пятница)
PROP 2: discover (board, PROP 3)	ПРОП-Я 2: понимать (совет, ПРОП-Я 3)
PROP 3: bankrupt (corporation)	ПРОП-Я 3: банкротство (корпорация)
PROP 4: when (PROP 1, PROP 2)	ПРОП-Я 4: когда (ПРОП-Я 1, ПРОП-Я 2)

Понимание второго предложения обеспечивается благодаря анафорической референциальной связи (referential cohesion) местоимения *they* и антецедента *board*. Очевидно, что ситуационная модель данного текста может содержать указания на «глубинные связи» (англ.deep cohesion), такие как увольнение президента советом, некомпетентность президента как причина банкротства, новый президент, платежеспособность корпорации. Coh-Metrix рассчитывает следующие типы *лексической кореферентности:* повторы (overlaps) нарицательных существительных, местоимений, основ и знаменательных частей речи как в смежных предложениях, так и во всем тексте. *Глубинная связность* на уровне предложений и текста рассчитывается на основе частоты встречаемости следующих типов дискурсивных маркеров: аддитивные (*также, кроме того*), темпоральные (*а затем, после, во время*), каузальные (*потому что, так*) и логические (*следовательно, если, и, или*). Частота вхождений каждого класса дискурсивных маркеров нормализуется на 1000 словоформ. *Лексическое разнообразие* рассчитывается как отношение неповторяющихся в тексте слов (англ. types) и словоформ (англ. tokens).

Важным условием понимания текста является способность моделирования *ситуационной (или ментальной) модели*, «референциального содержания или микромира того, о чем текст» (Graesser et al. 1994: 375). Ситуационная модель текста эксплицируется в причинно-следственных связях, интенциональности, темпоральности, пространственности и количестве субъектов коммуникативной ситуации (Zwaan & Radvansky 1998), рассчитываемых при помощи следующих метрик: количество каузальных глаголов, каузальных структур (*потому что, как следствие, как результат*), интенциональных глаголов, интенциональных структур (*чтобы, с помощью, посредством*), морфологических повторов (*времени и вида глагола*), отношение каузальных структур к каузальным глаголам, отношение интенциональных структур к интенциональным глаголам и др. Очевидно, например, что маркеры несовпадающих временных форм затрудняют построение ситуативной модели, а присутствие темпоральных маркеров (*позднее, до того, как, накануне*),

наоборот, снижают сложность текста. *Пространственность* как «пространственная плотность» (spatial density) и «пространственная связность» (spatial cohesion) оценивается путем подсчета доли существительных с семантикой местоположения и глаголов движения.

В дополнение к обсуждавшимся ранее предикторам кореференции, Coh-Metrix также оценивает *концептуальное сходство* между предложениями и абзацами при помощи латентного семантического анализа (LCA), статистического метода представления знаний о мире, основанного на идее семантического сходства слов, имеющих аналогичное окружение. На методе LCA основан еще один предиктор Coh-Metrix – *отношение объема заданной и новой* информации в тексте – рассчитываемый в двух вариантах: как среднее значение и стандартное отклонение (LSA given/new, sentences, mean, LSA given/new, sentences, standard deviation).

Разработчики Coh-Metrix нашли весьма изящное решение проблемы оценки сложности текстов различных жанров. Поскольку «разные типы текстов сложны по-разному», а формулы сложности жанрозависимы (см. Solnyshkina et al. 2020), определение жанровой принадлежности весьма затруднительно. В качестве причин укажем на отсутствие общепринятого набора текстовых категорий отдельных жанров и вероятностный характер присутствия в текстах одного жанра текстовых категорий нескольких жанров. Учитывая, что объекты «могут иметь категориальное отношение друг к другу только посредством обладания общими категориальными признаками» (см. Rosch & Mervis 1975: 603), можно было бы рассчитывать количество жанровых признаков, присутствующих в каждом конкретном тексте. При этом значимыми остаются два вопроса: нужно ли при оценке сложности оценивать присутствие категориальных признаков всех жанров в каждом конкретном тексте или достаточно выбрать, например, два–три жанра, имеющих стандартизированные критерии сложности и категориальные признаки которых можно аннотировать и рассчитать в большом корпусе. В качестве таких жанров разработчиками Coh-Metrix были выбраны нарративы (художественные тесты), учебные тексты по истории и естественнонаучные тексты, поскольку (1) обязательная программа школьного образования строится преимущественно на данных типах текстов и (2) именно для этих жанров были рассчитаны предикторы сложности (количество глаголов прошедшего времени, длина и частотность слов (McCarthy et al. 2009). Гипотеза, положенная в основу исследования, была следующая: категории текстов различных жанров можно ранжировать по сложности, а затем, определяя долю этой категории в тексте, определять их сложность. Например, если текст X содержит текстовую категорию C, и если категория C имеет уровень сложности D, то текст X унаследует уровень сложности D (McNamara et al. 2014: 13). Именно поэтому разработчики Coh-Metrix предлагают осуществлять классификацию текстов для оценки их сложности не по жанровой принадлежности, а по присутствию в них категорий того или иного жанра (McNamara et al.

2014: 5–6). Применение таких методик позволяет, например, выявить в тексте 70 % повествовательности и 30 % информативности.

Многочисленные исследования валидировали алгоритмы Coh-Metrix, доказав ее значимость как для исследовательских, так и практических целей: Coh-Metrix обеспечивает надежный анализ пяти уровней дискурса.

6. Заключение

Современная парадигма дискурсивной комплексологии как интегрального научного направления, предметом которого является оценка абсолютной и относительной сложности дискурса, сформирована на фундаменте лингвистических и психолингвистических достижений, а также успехов в области компьютерной лингвистики. Пройдя пять этапов своего развития – формирующего, классического, закрытых тестов, структурно-когнитивного и обработки естественного языка – дискурсивная комплексология разработала и валидировала более двухсот предикторов абсолютной сложности и десятки критериев относительной сложности. Дискурсивная комплексология подняла проблематику сложности текста на новый уровень, доказав, что оценка сложности текста должна быть дополнена оценкой когнитивных и лингвистических способностей языковой личности реципиента текста, а также анализом коммуникативной ситуации.

Специфика современного этапа – этапа искусственного интеллекта – состоит в использовании как традиционного «параметрического подхода», так и новых методов – методов машинного обучения для создания текстовых профайлеров, осуществляющих оценку сложности, подбор и модификацию текстов. Одним из наиболее успешных проектов в данной области следует признать разработку автоматического анализатора текстов на английском языке Coh-Metrix, в которой реализована пятиуровневая когнитивная модель восприятия текста. В дополнении к дескриптивным и «классическим» параметрам, таким как длина текста и индекс читабельности, Coh-Metrix успешно осуществляет оценку параметров текстового уровня, ситуационной модели и жанровых характеристик.

Перспектива научных исследований лингвистической комплексологии состоит в параметризации текстов – разработке перечня типологических параметров сложности текстов различных типов и жанров. Параметризация текстов разных языков позволит автоматизировать процесс подбора текстовых материалов для решения образовательных, социальных и профессиональных задач, а также добиться большей точности внутри- и межъязыковых исследований.

Благодарность

Работа выполнена за счет средств Программы стратегического академического лидерства Казанского (Приволжского) федерального университета (ПРИОРИТЕТ-2030).

Acknowledgments

This paper has been supported by the Kazan Federal University Strategic Academic Leadership Program (PRIORITY-2030).

REFERENCES / СПИСОК ЛИТЕРАТУРЫ

Anderson, Philip. 1972. More is different: Broken symmetry and the hierarchical nature of science. *Science* 177 (4047). 393–396.

Biber, Douglas. 1988. *Variation Across Speech and Writing*. Cambridge, England: Cambridge University Press. https://doi.org/10.1017/S0022226700014201

Biemiller, Andrew. 2009. *Words Worth Teaching*. Columbus, OH: SRA/McGraw-Hill.

Bormuth, John R. 1969. *Development of Readability Analysis*. Technical report, Projet number 7-0052, U.S. Office of Education, Bureau of Research, Department of Health, Education and Welfare, Washington, DC.

Bulté, Bram & Alex Housen. 2012. Defining and operationalising L2 complexity. In Housen Alex, Folkert Kuiken & Ineke Vedder (eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*, 21–46. Amsterdam: John Benjamins. https://doi.org/10.1075/lllt.32.02bul

Chall, Jeanne S. & Edgar Dale. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Cambridge: Brookline Books.

Charniak, Eugene. 2000. A maximum-entropyinspired parser. *In Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*. 132–139.

Coleman, Edmund B. 1965. *On Understanding Prose: Some Determiners of Its Complexity*. NSF Final Report GB2604, Washington, D.C, National Science Foundation.

Collins-Thompson, Kevyn. 2015. Computational assessment of text readability: A survey of current and future research. *ITL – International Journal of Applied Linguistics* 165 (2). 97–135.

Crossley, Scott A., Philip M. Mccarthy, David F Duffy & Danielle McNamara. 2007. Toward a new readability: A mixed model approach. *In Proceedings of the 29th Annual Conference of the Cognitive Science Society*. 197–202.

Dale, Edgar & Jeanne S. Chall. 1948. A formula for predicting readability. *Educational Research Bulletin* 27. 11–20, 37–54.

Dale, Edgar & Joseph O'Rourke. 1981. *Living Word Vocabulary*. Chicago: World Book – Childcraft International.

Danielson, Wayne A. & Sam D. Bryan. 1963. Computer automation of two readability formulas. *Journalism Quarterly* 40 (2). 201–205. https://doi.org/10.1177%2F107769906304000207

Daoust, François, Léo Laroche & Lise Ouellet. 1996. SATO-CALIBRAGE: Présentation d'un outil d'assistance au choix et à la rédaction de textes pour l'enseignement. *Revue Québécoise de Linguistique* 25 (1). 205–234.

Dascalu, Mihai. 2014. Analyzing discourse and text complexity for learning and collaborating. *In Analyzing Discourse and Text Complexity for Learning and Collaborating*, 1–3. Springer, Cham. https://doi.org/10.1007/978-3-319-03419-5

Flesch, Rudolf. 1948. A new readability yardstick. *Journal of Applied Psychology* 32 (3). 221–233. https://doi.org/10.1037/h0057532

Foltz, Peter W., Walter Kintsch & Thomas Landauer. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse Processes* 25 (2). 285–307. https://doi.org/10.1080/01638539809545029

Gatiyatullina, Galya, Marina Solnyshkina, Valery Solovyev, Andrey Danilov, Ekaterina Martynova & Iskander Yarmakeev. 2020. Computing Russian morphological distribution patterns using RusAC Online Server. *In 13th International Conference on Developments in eSystems Engineering (DeSE)*. 393–398. https://doi.org/10.1109/DeSE51703.2020.9450753

Graesser, Arthur C. & Danielle S. McNamara. 2011. Computational Analyses of Multilevel Discourse Comprehension. *Topics in Cognitive Science* 3. 371–398.

Graesser, Arthur C., Matthew Singer & Tom Trabasso. 1994. Constructing inferences during narrative text comprehension. *Psychological Review* 101. 371–395.

Gray, William & William Leary. 1935. *What Makes a Book Readable*. University of Chicago Press, Chicago: Illinois.

Hall, Charles, Debra S. Lee, Gwenyth Lewis, Phillip M. McCarthy & Danielle S. McNamara. 2006. Language in law: Using Coh-Metrix to assess differences between American and English/Welsh language varieties. *In Proceedings of the Annual Meeting of the Cognitive Science Society* 28.

Heilman, Michael, Le Zhao, Juan Pino & Maxine Eskenazi. 2008. Retrieval of reading materials for vocabulary and reading practice. *In Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*. 80–88. https://doi.org/10.3115/1631836.1631846

Hendrix, Gary G. 1980. Future prospects for computational linguistics. *In ACL '80: Proceedings of the 18th Annual Meeting on Association for Computational Linguistics*. 131–135. Association for Computational Linguistics, United States. https://doi.org/10.3115/981436.981476

Jones, Michael N., Walter Kintsch & Douglas J. Mewhort. 2006. High-dimensional semantic space accounts of priming. *Journal of Memory and Language* 55(4). 534–552.

Kemper, Susan. 1983. Measuring the inference load of a text. *Journal of Educational Psychology* 75 (3). 391–401.

Kintsch, Walter & Vipond Douglas. 1979. Reading comprehension and readability in educational practice and psychological theory. In Lars-Göran Nilsson (ed.), *Perspectives on memory research*, 329–365. Hillsdale, NJ, Lawrence Erlbaum.

Klare, George R. 1963. *The Measurement of Readability*. Iowa State University Press.

Kortmann, Bernd & Benedikt Szmrecsanyi (eds.). 2012. *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*. Berlin: De Gruyter.

Laposhina, Antonina N. & Maria Yu. Lebedeva. 2021. Tekstometr: Online-instrument opredeleniya urovnya slozhnosti teksta po russkomu yazyku kak inostrannomu. *Rusistika* 19(3). 331–345. (In Russ.) http://dx.doi.org/10.22363/2618-8163-2021-19-3-331-345

Lively, Bertha & Sidney Pressey. 1923. A method for measuring the 'vocabulary burden' of textbooks. *Educational Administration and Supervision* 9. 389–398.

Marujo, Luis, Jorge Baptista, José Lopes, Maxine Eskenazi, Ceu Viana, Juan Pino & Isabel Trancoso. 2009. Porting reap to European Portuguese. *In SLaTE*. 69–72. Citeseer.

McCall, William & Lelah Crabbs. 1925. *Standard Test Lessons in Reading*. New York: Teacher's College Press.

McCarthy, Philip M., John C. Myers, Stephen Briner & Arthur C. Graesser. 2009. A psychological and computational study of sub-sentential genre recognition. *JLCL* 24 (1). 23–55.

McClusky, Howard. 1934. A quantitative analysis of the difficulty of reading materials. *The Journal of Educational Research* 28. 276–282. https://doi.org/10.1080/00220671.1934.10880487

McLaughlin, G. Harry. 1969. Smog-grading – a new readability formula. *Journal of Reading* 13. 639–646.

McNamara, Danielle & Arthur C. Graesser. 2012. *Coh-Metrix: An Automated Tool for Theoretical and Applied Natural Language Processing*. IGI Global. https://doi.org/10.4018/978-1-60960-741-8.ch011

McNamara, Danielle S., Arthur C. Graesser, Philip M. McCarthy & Zhiqiang Cai. 2014. *Coh-Metrix: Theoretical, Technological, and Empirical Foundations. In Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511894664.006

Meyer, Bonnie J. F. 1982. Reading research and the composition teacher: The importance of plans. *College Composition and Communication* 33 (1). 37–49. https://doi.org/10.2307/357843

Nelson, Jessica, David Liben, Meredith Liben & Charles Perfetti. 2012. *Measures of Text Difficulty: Testing their Predictive Value for Grade Levels and Student Performance*. New York, NY: Student Achievement Partners.

Ojemann, Ralph. 1934. The reading ability of parents and factors associated with the reading difficulty of parent education materials. *University of Iowa Studies in Child Welfare* 8. 11–32.

Rabin, Mikhael'. 1993. Slozhnost' vychislenii. *In ACM Turing Award Lectures*. 371–391. Moscow: Mir. (In Russ.)

Rescher, Nicholas. 1998. *Complexity: A Philosophical Overview*. London: Transaction Publishers.

Rosch, Eleanor & Carolyn B. Mervis. 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology* 7. 573–605.

Rubakin, Nikolai A. 1890. Notes on literature for the people. *Russkoe Bogatstvo* 10. 221–231. (In Russ.)

Saimon, Gerbert. 2004. *The Sciences of the Artificial*. Moscow: Editorial URSS. (In Russ.)

Schwarm, Sarah E. & Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. *In ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. 523–530. https://doi.org/10.3115/1219840.1219905

Sheehan, Kathleen M., Irene Kostin, Diane Napolitano & Michael Flor. 2014. The TextEvaluator tool: Helping teachers and test developers select texts for use in instruction and assessment. *The Elementary School Journal* 115 (2). 184–209. https://doi.org/10.1086/678294

Sherman, Lucius A. 1893. *Analytics of Literature: A Manual for the Objective Study of English Prose and Poetry*. Boston: Ginn.

Si, Luo & Jamie Callan. 2001. A statistical model for scientific readability. *In Proceedings of the Tenth International Conference on Information and Knowledge Management*. 574–576. ACM New York, NY, USA. https://doi.org/10.1145/502585.502695

Simon, Herbert A. 1996. *The Sciences of the Artificial*. Cambridge: The MIT Press.

Smith, Edgar A. & John Quackenbush. 1960. Devereux teaching aids employed in presenting elementary mathematics in a special education setting. *Psychological Reports* 7. 333–336. https://doi.org/10.2466/PR0.7.6.333-336

Solnyshkina, Marina I., Elena V. Harkova & Aleksander S. Kiselnikov. 2014. Comparative Coh-metrix analysis of reading comprehension texts: Unified (Russian) state exam in English vs Cambridge first certificate in English. *English Language Teaching* 7 (12). 65–76. https://doi.org/10.5539/elt.v7n12p65

Solnyshkina, Marina I. & Kisel'nikov Aleksandr. S. 2015. Slozhnost' teksta: Etapy izucheniya v otechestvennom prikladnom yazykoznanii. *Vestnik Tomskogo Gosudarstvennogo Universiteta. Filologiya* 6(38). (In Russ.)

Solnyshkina, Marina I., Elena V. Harkova & Maria B. Kazachkova. 2020. The structure of Cross-Linguistic differences: Meaning and context of 'Readability' and its Russian

equivalent 'Chitabelnost'. *Journal of Language & Education* 6 (1). 103–119. https://jle.hse.ru/article/view/7176/12052. https://doi.org/10.17323/jle.2020.v6.i1

Solnyshkina, Marina I., Ehl'zara Gizzatullina-Gafiyatova, Ekaterina V. Martynova & Valery Solovyev. 2022. Text complexity as an interdisciplinary problem. *Voprosy Kognitivnoi Lingvistiki 1*. (In Russ.)

Solovyev, Valery D., Vladimir V. Ivanov & Marina I. Solnyshkina. 2018. Assessment of reading difficulty levels in Russian academic texts: Approaches and Metrics. *Journal of Intelligent & Fuzzy Systems* 34 (5). 3049–3058. https://doi.org/10.3233/JIFS-169489

Solovyev, Valery, Marina Solnyshkina, Vladimir Ivanov & Ildar Batyrshin. 2019. Prediction of reading difficulty in Russian academic texts. *Journal of Intelligent & Fuzzy Systems* 36 (5). 4553–4563. https://doi.org/10.3233/JIFS-179007

Solovyev, Valerii, Yulia Volskaya, Maria Andreeva & Artem Zaikin. 2022. Russian dictionary with concreteness/abstractness indexes. *Russian Journal of Linguistics* 2. 514–548. (In Russ.)

Spivey, Nancy N. 1987. Construing constructivism: Reading research in the United States. *Poetics* 16 (2). 169–192. https://doi.org/10.1016/0304-422X%2887%2990024-6

Steger, Maria & Edgar W. Schneider. 2012. Complexity as a function of iconicity: The case of complement clause constructions in New Englishes. In Kortmann Bernd & Benedikt Szmrecsanyi (eds.), *Linguistic complexity: Second language acquisition, indigenization, contact*, 156–191. Berlin: De Gruyter.

Stevens, Kathleen C. 1980. Readability Formulae and McCall-Crabbs Standard Test Lessons in Reading. *The Reading Teacher* 33 (4). 413–415.

Sun, Haimei. 2020. Unpacking reading text complexity: A dynamic language and content approach. *Studies in Applied Linguistics & TESOL at Teachers College* 20 (2). 1–20. https://doi.org/10.7916/salt.v20i2.7098

Taylor, Wilson L. 1953. Cloze procedure: A new tool for measuring readability. *Journalism Quarterly* 30 (4). 415–433. https://doi.org/10.1177%2F107769905303000401

Thorndike, Edward. 1921. Word knowledge in the elementary school. *The Teachers College Record* 22 (5). 334–370.

van Dijk, Teun A. & Walter Kintsch. 1983. *Strategies of Discourse Comprehension*. New York: Academic.

Vergara, Fermina & Rachelle Lintao. 2020. War on drugs: The readability and comprehensibility of illegal drug awareness campaign brochures. *International Journal of Language and Literary Studies* 2 (4). 98–121. https://doi.org/10.36892/ijlls.v2i4.412

Vogel, Mabel & Carleton Washburne. 1928. An objective method of determining grade placement of children's reading material. *The Elementary School Journal* 28 (5). 373–381. https://doi.org/10.1086/456072

Zwaan, Rolf A. & Gabriel A. Radvansky. 1998. Situation models in language comprehension and memory. *Psychological Bulletin* 123. 162–185. https://doi.org/10.1037/0033-2909.123.2.162

Zeno, Susan, Robert T. Millard & Raj Duvvuri. 1995. *The Educator's Word Frequency Guide*. Brewster: Touchstone Applied Science Associates, Inc.

Internet Resources / Электронные ресуры

Antonini, Alessio, Francesca Benatti, Edmund King, François Vignale & Guillaume Gravier. 2019. *Modelling Changes in Diaries, Correspondence and Authors' Libraries to Support Research on Reading: The READ-IT Approach*. URL: https://hal.archives-ouvertes.fr/hal-02130008/document (accessed 25 January 2022).

Antunes, Hélder M. M. 2019. *Automatic Assessment of Health Information Readability.* URL: https://repositorio-aberto.up.pt/bitstream/10216/121810/4/345408.pdf (accessed 25 January 2022).

Development of the ATOS Readability Formula. 2014. URL: https://webcache.googleusercontent.com/search?q=cache:lWV4zvGcnhMJ:https://doc.renlearn.com/KMNet/R004250827GJ11C4.pdf+&cd=14&hl=ru&ct=clnk&gl=ru (accessed 25 January 2022).

François, Thomas & Hubert Naets. 2011. Dmesure: A readability platform for French as a foreign language. URL: https://cental.uclouvain.be/team/tfrancois/articles/CLIN21.pdf (accessed 25 January 2022).

Lennon, Colleen & Hal Burdick. 2004. *The Lexile Framework as an Approach for Reading Measurement and Success*. URL: http://www.lexile.com/m/resources/materials/Lennon__Burdick_2004.pdf (accessed 25 January 2022).

Renaissance. 2022. URL: https://ukhosted43.renlearn.co.uk/2171850/ (accessed 25 January 2022).

Special Collections. Accelerated Reader (ATOS Level: 5.0-5.9). Bookshare a Benetech Initiative. 2002–2022. URL: https://www.bookshare.org/browse/collection/371895 (accessed 25 January 2022).

T.E.R.A.: The Coh-Metrix Common Core Text Ease and Readability Assessor. 2012–2022. URL: http://129.219.222.70:8084/Coh-Metrix.aspx (accessed 25 January 2022).

The ATOS Readability Formula for Books and How it Compares to Other Formulas. 2000. URL: https://files.eric.ed.gov/fulltext/ED449468.pdf (accessed 25 January 2022).

The Lexile Framework for Reading. 2022. URL: https://lexile.com (accessed 25 January 2022).

Article history:
Received: 20 October 2021
Accepted: 06 February 2022

Bionotes:

Marina I. SOLNYSHKINA is Doctor Habil. (Philology), Professor of the Department of Theory and Practice of Foreign Language Teaching, Head of "Text Analytics" Research Lab at the Institute of Philology and Intercultural Communication of Kazan Federal University (Russia). Her research interests include linguistic complexology, corpus linguistics, and lexicography.

Contact information:
Kazan (Volga Region) Federal University
18 Kremlevskaya str., Kazan, 420008, Russia
e-mail: mesoln@yandex.ru
ORCID: 0000-0003-1885-3039

Danielle S. MCNAMARA, Ph.D., is Professor of Psychology in the Psychology Department and Senior Scientist at Arizona State University. She is an international expert in the fields of cognitive science, comprehension, natural language processing, and intelligent systems.

Contact information:
Arizona State University
Payne Hall, TEMPE Campus, Suite 108, Mailcode 1104, the USA
e-mail: Danielle.McNamara@asu.edu
ORCID: 0000-0001-5869-1420

Radif R. ZAMALETDINOV is Doctor Habil. (Philology), Professor, Director of the Institute of Philology and Intercultural Communication of Kazan Federal University. His research interests embrace cognitive linguistics, linguoculturology, comparative linguistics, history and patterns of functioning of the Tatar and Russian languages, and bilingualism.

Contact information:
Kazan Federal University
18 Kremlevskaya str., Kazan, 420008, Russia
e-mail: director.ifmk@gmail.com
ORCID: 0000-0002-2692-1698

Сведения об авторах:
Марина Ивановна СОЛНЫШКИНА – доктор филологических наук, профессор кафедры теории и практики преподавания иностранных языков, руководитель НИЛ «Текстовая аналитика» Института филологии и межкультурной коммуникации Казанского федерального университета (Россия). Сфера ее научных интересов включает лингвистическую комплексологию, корпусную лингвистику и лексикографию.
Контактная информация:
Казанский (Приволжский) федеральный университет
Россия, 420008, Казань, ул. Кремлевская, д. 18
e-mail: mesoln@yandex.ru
ORCID: 0000-0003-1885-3039

Даниэль С. МАКНАМАРА – доктор наук, профессор кафедры психологии Университета штата Аризона, психолингвист, международный эксперт в области когнитивистики, понимания, обработки естественного языка и интеллектуальных систем.
Контактная информация:
Даниэль С. МакНамара
Университет штата Аризона
Пэйн Холл, Кампус TEMPE, ком. 108, 1104, США
e-mail: Danielle.McNamara@asu.edu
ORCID: 0000-0001-5869-1420

Радиф Рифкатович ЗАМАЛЕТДИНОВ – доктор филологических наук, профессор, директор Института филологии и межкультурной коммуникации Казанского федерального университета. В сферу его научных интересов входят когнитивная лингвистика, лингвокультурология, сопоставительное языкознание, история и закономерности функционирования татарского и русского языков, билингвизм.
Контактная информация:
Казанский (Приволжский) федеральный университет
Россия, 420008, Казань, ул. Кремлевская, д. 18
e-mail: director.ifmk@gmail.com
ORCID: 0000-0002-2692-1698