https://doi.org/10.17323/jle.2024.22221

Прогнозирование влияния многоуровневых лингвистических особенностей на читаемость учебников для начальной школы Гонконга: исследование, основанное на машинном обучении

Чжэнъе Сюй, Исюнь Ли, Дью Лю

Образовательный университет Гонконга , г. Тай По, Н.Т., Гонконг, Китай.

АННОТАЦИЯ

Введение: Формулы оценки читаемости играют важную роль в подборе подходящих текстов для развития навыков чтения у детей. Однако традиционные формулы, как правило, представляют собой линейные модели, разработанные для алфавитных языков, и испытывают трудности при учёте большого количества факторов.

Цель: Разработать новые формулы оценки читаемости китайских текстов с использованием алгоритмов машинного обучения, способных обрабатывать сотни факторов. Это первая формула читаемости, созданная для китайских текстов в Гонконге.

Методология: В исследовании использовался корпус из 723 текстов, взятых из 72 учебников китайского языка, применяемых в начальных школах. В качестве факторов рассматривались 274 лингвистических признака на уровне иероглифов, слов, синтаксиса и дискурса. В качестве зависимых переменных использовались уровень сложности текста по шкале, назначенной издателями, и оценка читаемости, выставленная учителями. Были обучены 15 моделей, основанных на различных комбинациях лингвистических признаков, с применением алгоритмов опорных векторов (SVM) и случайного леса (RF). Эффективность моделей оценивалась по точности предсказания и среднему абсолютному отклонению между предсказанным и реальным уровнем читаемости. Для обеих зависимых переменных наилучшие результаты показали модели на основе всех уровней признаков и модели, использующие только признаки на уровне иероглифов, построенные с помощью алгоритма случайного леса. Также был проведён анализ 10 наиболее значимых признаков в этих оптимальных моделях.

Результаты: Среди всех моделей наилучшие результаты показали модели на основе алгоритма случайного леса (Random Forest), использующие все уровни признаков (all-RF) и только признаки на уровне иероглифов (character-RF). Анализ значимости признаков в этих двух оптимальных моделях показал, что на оценку читаемости китайских текстов в контексте Гонконга сильнее всего влияют такие факторы, как последовательность изучения иероглифов, частотность иероглифов и частотность слов. Кроме того, результаты указывают на то, что издательства, вероятно, используют более широкий спектр источников информации при назначении уровня сложности текста (семестра), в то время как учителя, по-видимому, предпочитают ориентироваться на показатели, которые можно непосредственно извлечь из самого текста.

Заключение: Исследование подчёркивает важность признаков на уровне иероглифов, особенно того, когда иероглиф впервые появляется в учебнике, при прогнозировании читаемости китайских текстов в условиях Гонконга.

КЛЮЧЕВЫЕ СЛОВА

китайский язык, лингвистические признаки, метод случайный лес (алгоритм машинного обучения), модели читаемости, метод опорных векторов

Для цитирования: Сюй, Ч., Ли, И., & Лю, Д. (2024). Прогнозы влияния многоуровневых лингвистических особенностей на читаемость учебников начальной школы в Гонконге: Исследование на основе машинного обучения. *Journal of Language and Education, 10*(4), 151-163. https://doi.org/10.17323/jle.2024.22221

Correspondence:
Дью Лю,
duoliu@eduhk.hk

Получена: 14 августа 2024

Принята: 16 декабря 2024

Опубликована: 30 декабря 2024

ВВЕДЕНИЕ

Читаемость текста — это степень легкости, с которой текст может быть прочитан и понят (Crossley и др., 2019). Ряд исследований на разных языках показал, что на читаемость влияют как характеристики самого читателя — например, его языковая компетентность и мотивация (Stutz и др., 2016; Zhang и др., 2014), — так и особенности самого текста, такие как частота слов и длина предложений. Настоящее исследование сосредоточено на изучении текстов из китайских учебников для начальных школ в Гонконге с целью улучшения соответствия между уровнем сложности текстов и уровнем подготовки школьников, что, в свою очередь, способствует более эффективному обучению китайскому языку.

Читаемость текста и лингвистические признаки разных уровней. Читаемость текста может быть количественно оценена с помощью специальных формул (Crossley и др., 2019), которые позволяют определить уровень сложности текста. Такие формулы полезны для более точной оценки способности детей к восприятию текста и помогают подбирать материалы, соответствующие их уровню. Обычно формулы читаемости выдают абсолютное значение или уровень, соответствующий школьному классу, для которого текст считается подходящим (Kincaid и др., 1975; Solnyshkina и др., 2017). Например, одна из самых известных — формула Флеша–Кинкаида (Flesch-Kincaid), рассчитывается по следующей формуле: (0.39 × среднее количество слов в предложении) + (11.8 × среднее количество слогов в слове) – 15.59 Полученное значение указывает на рекомендуемый класс. Например, результат 5.3 означает, что текст подходит для учеников пятого класса. Однако такие формулы обычно используют ограниченное число признаков, несмотря на то что исследования показывают: характеристики на уровне слов, синтаксиса и дискурса оказывают значительное влияние на понимание текста на разных языках, включая английский и китайский (Crossley и др., 2019; Liu и др., 2024; Pinney и др., 2024; Solnyshkina и др., 2017). На лексическом уровне ключевым показателем читаемости является длина слова (количество иероглифов). Длинные слова, как правило, делают текст сложнее, тогда как короткие — наоборот, упрощают восприятие (Crossley и др., 2023; Mesmer & Hiebert, 2015). Варьирование словарного состава, то есть разнообразие слов, также влияет на читаемость (Sung и др., 2015). Частота употребления слов и психолингвистические показатели (например, время реакции и уровень ошибок в лексических тестах) также связаны с уровнем читаемости (Tsang и др., 2018; Tse и др., 2017). Помимо узнавания слов, важно также знание их значений — семантическая составляющая текста имеет решающее значение для его понимания (Mesmer & Hiebert, 2015). Грамматическая категория слов (например, существительные, глаголы, наречия) также влияет на читаемость. Тексты с высоким уровнем читаемости чаще содержат союзы и наречия, в то время как в менее читаемых текстах преобладают прилагательные и модальные слова (Liu и др., 2024). На синтаксическом уровне значимым фактором является длина предложения: более длинные предложения и большие расстояния между связанными словами указывают на более высокую синтаксическую сложность (Crossley и др., 2023). Например, зависимость слов (то есть среднее расстояние между двумя связанными словами) оказывает влияние на сложность восприятия. Чем ближе связаны слова друг к другу — тем легче текст воспринимается (Crossley и др., 2019). Также важны грамматические структуры предложений, определяющие логические связи между словами (Graesser и др., 2011). На дискурсивном уровне большое значение имеют связи между предложениями. Структура текста, обеспечивающая его связность, помогает читателю понять, как разные части текста соотносятся друг с другом (Pinney и др., 2024). Например, причинно-следственные связи, выраженные с помощью союзов и других логических маркеров, могут как облегчать, так и усложнять восприятие текста, в зависимости от их плотности и ясности (Graesser и др., 2011; Givón, 1995). Хотя характеристики самого читателя уже достаточно хорошо изучены в контексте чтения (McBride-Chang и др., 2005; Stutz и др., 2016), особенности самих текстов, особенно на китайском языке, до сих пор получили меньше внимания (Crossley и др., 2023; Fitzgerald и др., 2015; Sung и др., 2015). Настоящее исследование направлено на восполнение этого пробела — оно рассматривает влияние лингвистических признаков текста на его читаемость в контексте китайского языка в начальной школе Гонконга.

Исследование читаемости текста на китайском языке

Исследование читаемости текста на китайском языке включает в себя особенности на уровне слов, синтаксиса и дискурса, как в алфавитных языках, но также учитывает особенности на уровне символов, поскольку иероглиф является основной единицей письма в китайском языке (Cheng et al., 2020; Sung et al., 2015). В частности, символ может стоять отдельно, образуя односимвольное слово (например, 筆/bat1/ручка) или могут быть объединены с другими для образования двухсимвольных слов (например, 筆記/bat1-gei3/ примечание) или слов из трех и более символов (например, 筆記本/bat1-gei3-bun2/ блокнот). Каждые китайские иероглифы имеют свою собственную форму, звучание и значение (значения); следовательно, связанные с иероглифами лингвистические особенности могут влиять на читаемость текста (Sung et al., 2015). Традиционные формулы удобочитаемости китайского текста включают характеристики на уровне символов, такие как среднее количество символов, но часто упускают из виду другие важные факторы, особенно на уровне дискурса, такие как связность текста (Cheng et al., 2020; Jing, 1995). Они также предполагают линейную зависимость между уровнем читаемости и лингвистическими особенностями, что ограничивает точность модели (Rodriguez-Galiano et al., 2015).

Для решения этих проблем были использованы методы машинного обучения, позволяющие улучшить оценку читаемости. В отличие от традиционных формул, машинное обучение может обрабатывать большое количество лингвистических характеристик и выявлять сложные взаимосвязи между ними (Rodriguez-Galiano et al., 2015). Этот подход представляет прогнозируемый уровень читаемости в виде категории (например, 5-й класс), что указывает на соответствующий уровень чтения для читателей в этом классе. Помимо содействия сопоставлению текста с текстом для чтения, подход машинного обучения может улучшить наше понимание читаемости текста, определяя ключевые лингвистические особенности (Родригес-Гальяно и др., 2015). Например, Фитцджеральд и соавторы (2015) проанализировали 238 признаков и определили, что девять признаков, связанных со структурой слова, семантикой и связностью, имеют решающее значение для понимания сложности английского текста. Таким образом, в текущем исследовании использовались подходы машинного обучения, чтобы получить представление о читабельности текста.

Машинное обучение и формулы оценки читаемости китайских текстов

Методы машинного обучения активно применяются для оценки читаемости китайских текстов. В большинстве исследований используется алгоритм опорных векторов (SVM). Так, например, работы Chen и др. (2011), Sung и др. (2015), Wu и др. (2020), проведённые преимущественно на Тайване, использовали традиционную китайскую письменность и признаки с нескольких лингвистических уровней для обучения моделей SVM, успешно классифицирующих тексты по уровню сложности. В частности, модели показали высокую точность при классификации текстов для младших классов (1–2 классы — 95%) и средних классов (3–4 классы — 84%; Chen и др., 2013). Для более точной градации читаемости по уровням классов (1–6 классы), Sung и др. (2015) применили SVM в сочетании с 31 лингвистическим признаком, охватывающим лексический, семантический, синтаксический и дискурсивный уровни. Результаты показали, что использование признаков с разных уровней значительно повышает точность модели (до 71,75%) по сравнению с использованием признаков только с одного уровня (43,97%–65,13%). Исследования читаемости текста на упрощённом китайском, используемом в материковом Китае, также активно проводятся. Wu и др. (2020) использовали SVM-модели для анализа влияния 104 лингвистических признаков на читаемость, включая признаки на уровне иероглифов, слов (состоящих из двух и более иероглифов), синтаксиса и дискурса. Их исследование показало, что среди моделей с признаками одного уровня наилучшие результаты дали признаки на уровне слов (точность: 62,1%). Тем не менее, добавление признаков на уровне иероглифов (63,8%) и синтаксиса (63,1%) улучшило точность предсказаний. Недавнее исследование Liu и др. (2024) изучало влияние лингвистических признаков на читаемость текстов на упрощённом китайском с использованием более точной шкалы сложности по семестрам (от 1 до 12). В исследовании применялись алгоритмы случайного леса (RF) и SVM, а также большое количество лексических и дискурсивных признаков. Результаты подтвердили, что модели, использующие признаки с разных уровней, превосходят модели с признаками только одного уровня: точность предсказаний составила 27% для RF и 28% для SVM, при этом средняя абсолютная ошибка (MAE) была ниже (RF: 1,24; SVM: 1,25). Кроме того, исследование выявило, что наиболее важными признаками являются частотность иероглифов и слов, семантические признаки, лексическое разнообразие, синтаксические категории и референтная связность.

Настоящее исследование

В отличие от Тайваня и материкового Китая, вопросам читаемости текстов в Гонконге уделялось значительно меньше внимания, несмотря на то, что китайский язык в этом регионе обладает уникальными особенностями. В Гонконге используется традиционная письменность, а устная речь основана на кантонском диалекте, который заметно отличается от используемых языковых вариантов в материковом Китае и на Тайване (McBride-Chang и др., 2005). Кантонский отличается рядом особенностей: наличием дополнительных тонов и уникальной лексики, употреблением специфических разговорных форм (например, «係/hai6/» вместо литературного «是/si6/» для выражения утвердительного ответа), использованием частиц в конце высказываний («啊/aa3/», «㗎/gaa3/», «囉/lo1/»), которые редко встречаются в формальных текстах, региональными различиями в лексике.

Так, автобус и такси в Гонконге называются «巴士/baa1-si6/» и «的士/dik1-si6/», тогда как в материковом Китае — «公交/gong1-jiao1/» и «出租车/chu1-zu1-che1», а на Тайване — «公車/gong1-che1/» и «計程車/ji4-cheng2-che1». Эти языковые особенности требуют создания формул читаемости, основанных на текстах, используемых именно в Гонконге. Настоящее исследование использует корпус текстов из учебников по китайскому языку для начальной школы, которые активно применяются в государственных школах Гонконга. Вслед за предыдущими исследованиями (например, Liu и др., 2024; Sung и др., 2015), в данном исследовании учитываются лингвистические признаки на уровне иероглифов, слов, синтаксиса и дискурса. Для оценки читаемости используется более точная шкала — от 1 до 24 (по семестрам обучения), а также субъективный показатель — оценки читаемости, выставленные учителями для каждого текста.В качестве алгоритмов были выбраны SVM и случайный лес (RF). Также проведён анализ значимости признаков для лучшего понимания факторов, влияющих на читаемость китайских текстов.Цель исследования — ответить на два ключевых вопроса: Влияют ли уровни лингвистических признаков на эффективность моделей оценки читаемости и в какой степени? Какие признаки являются наиболее значимыми в лучших моделях?

МЕТОДОЛОГИЯ

Дизайн исследования

В исследовании использовался корпус текстов из учебников по китайскому языку для начальной школы, изданных тремя крупнейшими издательствами Гонконга. В связи с вопросами авторского права текстовые материалы не подлежат публичному распространению. Каждое издательство предоставило по четыре учебника на каждый уровень — по два на семестр, что в общей сложности составило 72 учебника. Два научных ассистента тщательно оцифровали и трижды вычитали тексты для обеспечения точности. В анализ были включены 723 текста, за исключением неполноценных фрагментов, таких как классическая китайская проза, иллюстрации, оглавления, библиографии и указатели. Далее из каждого текста были извлечены и рассчитаны лингвистические признаки, отражающие характеристики на уровнях иероглифа, слова, синтаксиса и дискурса, с использованием системы китайской сегментации CKIP (Ma & Chen, 2005).

Исследование было одобрено Комитетом по этике исследований на людях Образовательного университета Гонконга и проводилось в соответствии с Хельсинкской декларацией.

Машинное обучение для анализа читаемости осуществлялось с помощью библиотеки scikit-learn версии 1.1.2 на языке Python 3.10 (Pedregosa и др., 2011). В качестве зависимых переменных (Y) использовались два показателя читаемости текста:

- Y1 — семестр, назначенный издателем,
- Y2 — семестр, оценённый учителями.

Оценку от учителей составляли средние значения читаемости текста, выставленные 11 опытными учителями начальной школы (каждый из которых предоставил письменное информированное согласие на участие), по 9-балльной шкале, адаптированной под читателя среднего уровня для каждого класса. На основе усреднённых оценок в каждом классе текстам были присвоены значения по шкале от 1 до 24, где более высокие значения означали более сложные тексты. Были сформированы 15 различных комбинаций лингвистических признаков (обозначенных как X), включающих: признаки одного уровня: C (иероглифы), W (слова), S (синтаксис), D (дискурс), признаки двух уровней: C_W, C_S, C_D, W_S, W_D, S_D, признаки трёх уровней: C_W_S, C_W_D, C_S_D, W_S_D, признаки всех четырёх уровней: Xall.

В соответствии с подходами Liu и др. (2024) и Sung и др. (2015), были применены два алгоритма машинного обучения — метод опорных векторов (SVM) и случайный лес (RF). Для оценки эффективности моделей использовалась процедура перекрёстной проверки на 5 подвыборках. Все 723 текста были случайным образом разбиты на пять равных по количеству семестров подгрупп. В каждом прогоне четыре подгруппы использовались для обучения модели, а одна — для тестирования. Предсказанные значения Y сравнивались с фактическими значениями для оценки точности и средней абсолютной ошибки (MAE). Для статистического сравнения моделей читаемости были построены линейные смешанные модели (LMMs) с использованием пакета lmer в R 4.0.3 (Baayen и др., 2008). Для устранения мультиколлинеарности данные были стандартизированы (z-преобразование), а в модель были включены: тип признаков X, алгоритм машинного обучения и их взаимодействие как фиксированные эффекты. В качестве базовых уровней использовались RF и Xall. Также учитывались случайные перехваты и наклоны, а более сложная модель принималась, если улучшалась аппроксимация данных. После сравнения моделей были выбраны лучшие модели для предсказания семестров, присвоенных издателями и учителями. Затем была определена значимость каждого признака с использованием метода перестановочной важности признаков в пакете ELI5 для Python. Суть метода заключается в оценке потери точности предсказания и увеличения MAE при исключении отдельного признака из модели (Korobov & Lopuhin, 2019). Данные и исходный код анализа открыто размещены в репозитории Open Science Framework.

Лингвистические характеристики

Характеристики на уровне символов

Всего 110 характеристик на уровне символов, относящихся к четырем аспектам: 1) разнообразие символов (N = 4), 2) структурная сложность символов (N = 8), 3) частота встречаемости символов (N = 40) и 4) для каждого текста была рассчитана психолингвистическая информация о персонажах (N = 58). Для определения разнообразия символов учитывались четыре показателя: количество исходных символов в токене (для всех символов) , количество исходных типов (для разных символов), отношение количества типов к количеству токенов и доля символов которые встречаются только один раз. Например, в одном из текстов, «我和妈妈玩捉迷藏», есть восемь условных знаков. Поскольку третий и четвертый символы в предложении совпадают, то есть 妈, существует семь типов символов. Структурная сложность символов измерялась с помощью четырех показателей: среднего количества штрихов, доли символов с менее чем десятью штрихами, от 10 до 20 штрихов и более чем 20 штрихов. Количество лексем и количество типов было рассчитано для каждого показателя, в результате чего было получено восемь лексических характеристик структурной сложности символов. Частота встречаемости символов была измерена с использованием пяти корпусов: Сбалансированного корпуса современного китайского языка, CNCORPUS (Jin et al., 2005), The SUBTLEX-CH corpus (Cai & Brysbaert, 2010), Sinica Corpus (Huang, 200 6), Chinese text computing (Da, 2004) и «Гонконг, материковый Китай и Тайвань: частотность китайских иероглифов» (далее этот корпус будет обозначен как HKMCT 2). Симво-

лы, которые не были найдены в данном корпусе, считались сложными символами. Учитывались четыре показателя: средняя частота встречаемости часто встречающихся символов, стандартное отклонение частотности часто встречающихся символов, часто используемые символы, общее количество сложных символов и доля сложных символов. Для расчета этих показателей использовались количество токенов и количество типов символов.

Психолингвистическая информация о персонажах оценивалась с использованием 26 показателей из предыдущих исследований (например, Liu et al., 2007; Su et al., 2023). Этими показателями были: возраст, в котором, как ожидается, можно будет выучить тот или иной персонаж, знакомство с персонажем, легкость описания значения персонажа, легкость создания образа персонажа, класс и семестр, в котором персонаж впервые появился в учебнике, количество значений символа, число количество омофонов символа, суммарная частота всех символов, которые имеют одинаковое произношение (рассчитывается на основе пяти вышеупомянутых корпусов), количество слов, которые может образовывать символ, суммарная частота всех слов, содержащих символ (рассчитывается на основе пяти вышеупомянутых корпусов), наличие признаки произношения иероглифов (1 - достоверный признак, 0 - отсутствие признаков и -1 - ненадежный/вводящий в заблуждение), а также время реакции и частота ошибок при назывании иероглифов взрослыми китайцами.

Два показателя, касающихся подсказок произношения в символах были включены: доля символов с надёжным. подсказки произношения (подсказка произношения имеет то же. произношение, что и символ) и те, у которых нет/. вводящие в заблуждение подсказки произношения (звуки подсказки произношения и его соответствующий характер отличаются; Su. et al., 2023). Семантическая радикальная прозрачность символов, которая относится к степени соответствия значения между семантическим радикалом и всем символом, также была вовлечена. Например, в то время как 海/hoi2/ (sea) и 测/. cak1/ (мера) содержит семантический радикал 氵 (вариант. 水/seoi2/ воды), первый - семантически прозрачный, а второй - непрозрачный. Количество маркеров и количество типов было рассчитано для каждого показателя, что привело к появлению 58 лексических признаков для этой категории.

Функции на уровне слов

Для каждого текста было рассчитано в общей сложности 105 характеристик на уровне слов , охватывающих шесть аспектов: длина слова (N =12), разнообразие слов (N = 4), частота встречаемости слов (N = 32), психолингвистическая информация о словах (N = 22), структура набора (N = 1), и частично - синтаксические категории речи (N = 34). Что касается длины слова, то учитывались шесть параметров: средняя длина слова и процентное содержание односимвольных, двухсимвольных, трехсимвольных, четырехсимвольных и пяти- и более символьных слов в тексте. Количество токенов и типов было рассчитано для каждого аспекта, в результате чего было получено 12 лингвистических характеристик. Разнообразие слов относится к богатству слов в тексте. Были рассмотрены четыре показателя: общее количество слов (как количество лексем, так и количество типов), отношение количества типов к количеству лексем и доля слов, которые встречаются только один раз. Частота употребления слова была измерена на основе частоты употребления слова в пяти корпусах, за исключением ГМКИ, поскольку у него нет статистики по частоте употребления слова. Были рассмотрены четыре показателя: средние значения частотности для часто встречающихся слов, стандартное отклонение значений частотности для часто встречающихся слов, общее количество сложных слов и доля сложных слов. Для расчета этих показателей использовались количество лексем и количество типов, в результате чего было получено 32 лексических признака. Психолингвистическая информация о словах была рассчитана на основе corpus MELD-SCH (Tsang et al., 2018), который предоставил время реакции и частоту ошибок для взрослых китайцев. Двенадцать признаков были рассчитаны на основе среднего значения и стандартного отклонения времени реакции и частоты ошибок. Слова, которые небыли включены в MELD-SCH, считались низкочастотными словами, которые были идентифицированы по их пропорциям и исходным числам. Семантическая радикальная прозрачность слов, состоящих из двух символов , была получена на основе работы Су и др. (2023). Для каждого символа и для слово в целом. Для всех психолингвистических характеристик было рассчитано количество лексем и количество типов для каждого аспекта. Структура набора была измерена путем вычисления исходного количества именованных объектов (например, имен людей, организаций и местоположений) в тексте с использованием HanLP3. Синтаксические категории частей речи были определены путем присвоения одной из 16 категорий каждому слову с использованием естественного языка. Платформа для обработки и обмена информацией (Liu et al., 2004). 16 категорий включают девять типов контента слова (существительные, глаголы, прилагательные, числительные, квантификаторы, местоимения, слова времени, слова места и позиционные слова) и семь типов служебных слов (наречия, предлоги, союзы, частицы,междометия, дифференцирующие слова и слова состояния). Было подсчитано исходное количество и пропорция слов для каждой категории, в результате чего было получено 32 признака. Кроме того, было подсчитано количество и пропорция всех слов контента , в результате чего было получено 34 признака дискурса.

Функции синтаксического уровня

На синтаксическом уровне было рассмотрено 15 функций, основное внимание уделялось длине предложения (N = 4), зависимости от слова (N = 3) и грамматике (N = 8). Длина предложения определялась четырьмя параметрами: среднее количество символов и слов рассчитывалось отдельно

для каждого предложения и каждого придаточного предложения. Отдельные предложения и придаточные предложения были определены на основе знаков препинания. Предложение заканчивается точкой, восклицательным знаком, вопросительным знаком, многоточием или тире, в то время как предложение заканчивается запятой, двоеточием или точкой с запятой (Wang & Wu, 2020). Зависимость слов была проанализирована в каждом предложении. В каждом предложении учитывались три показателя: количество символов и слов перед основным глаголом и среднее расстояние между словами между любыми парами связанных слов. Связанные слова относятся к словам, которые синтаксически подчинены или зависят от другого слова. Были рассмотрены четыре грамматических показателя, отражающих наличие сложной грамматики китайского языка: отрицательные, метафорические, страдательные и противопоставительные предложения, основанные на Открытая платформа Baidu(https://cloud.baidu.com). Мы подсчитали количество и пропорцию каждого из этих четырех типов предложений в тексте и, таким образом, выявили восемь грамматических особенностей дискурса.

Признаки на уровне дискурса

На дискурсивном уровне были проанализированы 24 признака референтной связности и 20 признаков причинной связности.

Референтная связность

Эти признаки были включены на основе работы Graesser и др. (2011) и отражают, насколько элементы текста связаны между собой с помощью повторяющихся слов. Анализировались четыре типа слов: все слова, содержательные слова (content words), существительные, глаголы. Для каждого из этих типов рассчитывались шесть показателей: Доля соседних пар предложений или абзацев, содержащих одинаковые слова. Доля всех возможных пар предложений или абзацев, содержащих одинаковые слова. Взвешенная доля всех возможных пар предложений или абзацев, содержащих одинаковые слова — с учётом расстояния между предложениями/абзацами. Вес рассчитывался по формуле 1/(L + 1), где L — количество предложений между сравниваемыми единицами. Все эти расчёты выполнялись отдельно для предложений и для абзацев.

Причинная связность

Признаки причинной связности также были заимствованы из Graesser и др. (2011) и отражали наличие логических связей между частями текста. Рассчитывались абсолютные количества слов-связок, выражающих различные типы логических отношений: указание на предшествование (например, сначала), причинно-следственные связи (потому что), противопоставления (однако), сочинения (и), добавления (более того), последовательности (затем), умозаключения (только если), условия (если не), предположения (если), уступки (хотя), цели (для того чтобы), частотности (всегда), вставных конструкций (как всем известно), отказа или прекращения действия (лучше не), результата (поэтому), сравнений (а не), предпочтений (вместо того чтобы),обобщений (в целом), примеров (например), времени (когда). Эти лексические элементы обеспечивают логическую связность текста и тем самым влияют на его читаемость.

РЕЗУЛЬТАТЫ

Роль лингвистических особенностей в прогнозированиичитабельность текста на китайском языке

Средние значения и стандартные отклонения точности и MAE пятикратной перекрестной проверки приведены в таблице 1. Результаты моделей Y1 и Y2 были аналогичными (см. рисунок 1). Во всех четырех LMM наблюдался значительный эффект от применения алгоритма машинного обучения. RF превзошел SVM в прогнозировании удобочитаемости текста с более высокой точностью (Y1: Оценка = -1,27, SE = 0,25, t(145) = -5,08, p < 0,001; Y2: Оценка = -1,38, SE = 0,24, t(145) = -5,87, p < 0,001) и ниже среднего значения (Y1: Оценка = 0,76, SE = 0,19, t(139,99) = 3,90, p < 0,001; Y2: Оценка = 0,56, SE = 0,18, t(145) = 3,08, p = .003). Что касается X, то Xall продемонстрировал превосходную производительность по сравнению с Xs без функций символьного уровня, за исключением W_S_D вМодели Y2 (ps > .05). Это было очевидно с точки зрения точности (Оценки = -1,25 – -3,03, SEs = 0,24 – 0,25, ts = -12,86 – -4,99, ps < .001) и MAE (оценки = 0,50 – 2,72, SEs = 0,18, ts = 2.76 – 15.21, ps < .01). Однако не было обнаружено существенных различий между Xall и Xs, которые включали бы характеристики на уровне персонажа (ps > .05).

Взаимодействия между алгоритмом машинного обучения и контрастами между Xall и Xs без признаков символьного уровня, за исключением W_S_D, были значительными в моделях точности (оценки = 1,03 - 1,63, SEs = 0,33 – 0,35, ts= 2,91 – 4,88, ps < .01). Что касается MAE, то взаимодействия между алгоритмом машинного обучения и контрастами Xall с S (оценка = -0,52, SE = 0,26, t = -2,01, p = 0,047) и D (оценка = 0,61, SE = 0,26, t = 2,39, p = 0,019) былизначительными, в то время как другие взаимодействия были незначительными. были незначительными (ps > .05). Последующий анализ показал, что среди моделей RFразличия между C и Xall как с точки зрения точности, так и с точки зрения MAE были незначительными (ps > .05). Более того, C показал более высокую точность, чем Xs, которые не учитывали характерные особенности (Y1: Оценки = 1,12 – 2,66, SEs = 0,28, ts = 3,99 – 9,48, ps < .05; Y2: Оценки = 1,20 – 2,72, SEs = 0,26, ts = 4,56 – 10,32, ps < .01). Кроме того, у C был более низкий MAE, чем у Xs без символов и слов (т.е. S, D и S_D, Y1: оценки = -2,74 – -2,06, SE = 0,20, ps < .001; Y2: Оценки = -2,43 – -1,66, SE = 0,20, ts= -8,20 – -11,98, ps < .01).

Таблица 1

Средние значения и стандартные отклонения точности (ACC) и средней абсолютной ошибки (MAE) всех моделей машинного обучения

		Y1 (Семестр, назначенный издателем)		Y2 (Семестр с оценкой преподавателя)	
	X	**ACC**	**MAE**	**ACC**	**MAE**
SVM	All	0.21 (0.02)	2.45 (0.18)	0.21 (0.02)	2.45 (0.14)
	C	0.20 (0.02)	2.24 (0.09)	0.20 (0.02)	2.29 (0.14)
	W	0.20 (0.03)	2.63 (0.23)	0.20 (0.02)	2.62 (0.20)
	S	0.14 (0.03)	3.59 (0.33)	0.13 (0.03)	3.55 (0.27)
	D	0.12 (0.02)	4.23 (0.53)	0.12 (0.02)	4.32 (0.54)
	C_W	0.21 (0.04)	2.34 (0.27)	0.21 (0.03)	2.33 (0.22)
	C_S	0.20 (0.02)	2.35 (0.08)	0.20 (0.02)	2.40 (0.11)
	C_D	0.20 (0.03)	2.46 (0.21)	0.20 (0.03)	2.47 (0.20)
	W_S	0.20 (0.03)	2.68 (0.22)	0.20 (0.03)	2.66 (0.22)
	W_D	0.19 (0.04)	2.81 (0.18)	0.19 (0.03)	2.86 (0.12)
	S_D	0.15 (0.02)	3.66 (0.37)	0.15 (0.01)	3.78 (0.36)
	C_W_S	0.20 (0.02)	2.41 (0.18)	0.20 (0.02)	2.43 (0.15)
	C_W_D	0.21 (0.02)	2.43 (0.20)	0.22 (0.02)	2.40 (0.12)
	C_S_D	0.20 (0.03)	2.49 (0.10)	0.20 (0.03)	2.52 (0.07)
	W_S_D	0.21 (0.03)	2.78 (0.18)	0.22 (0.02)	2.40 (0.12)
RF	All	0.29 (0.04)	1.97 (0.13)	0.30 (0.02)	2.10 (0.19)
	C	0.28 (0.05)	1.96 (0.22)	0.28 (0.04)	2.10 (0.11)
	W	0.18 (0.04)	2.40 (0.04)	0.21 (0.02)	2.49 (0.08)
	S	0.15 (0.03)	3.37 (0.18)	0.15 (0.03)	3.52 (0.25)
	D	0.11 (0.02)	3.69 (0.20)	0.11 (0.02)	3.60 (0.11)
	C_W	0.29 (0.04)	2.02 (0.20)	0.29 (0.03)	2.06 (0.18)
	C_S	0.30 (0.04)	1.93 (0.15)	0.28 (0.02)	2.03 (0.10)
	C_D	0.29 (0.04)	2.02 (0.19)	0.28 (0.03)	2.08 (0.21)
	W_S	0.19 (0.03)	2.38 (0.13)	0.19 (0.04)	2.41 (0.19)
	W_D	0.21 (0.04)	2.37 (0.07)	0.20 (0.05)	2.51 (0.17)
	S_D	0.14 (0.02)	3.26 (0.09)	0.14 (0.03)	3.12 (0.28)
	C_W_S	0.28 (0.04)	1.96 (0.17)	0.30 (0.03)	2.05 (0.10)
	C_W_D	0.30 (0.04)	1.96 (0.15)	0.28 (0.04)	2.08 (0.14)
	C_S_D	0.30 (0.05)	1.95 (0.18)	0.30 (0.05)	2.07 (0.22)
	W_S_D	0.21 (0.04)	2.31 (0.14)	0.29 (0.03)	2.06 (0.13)

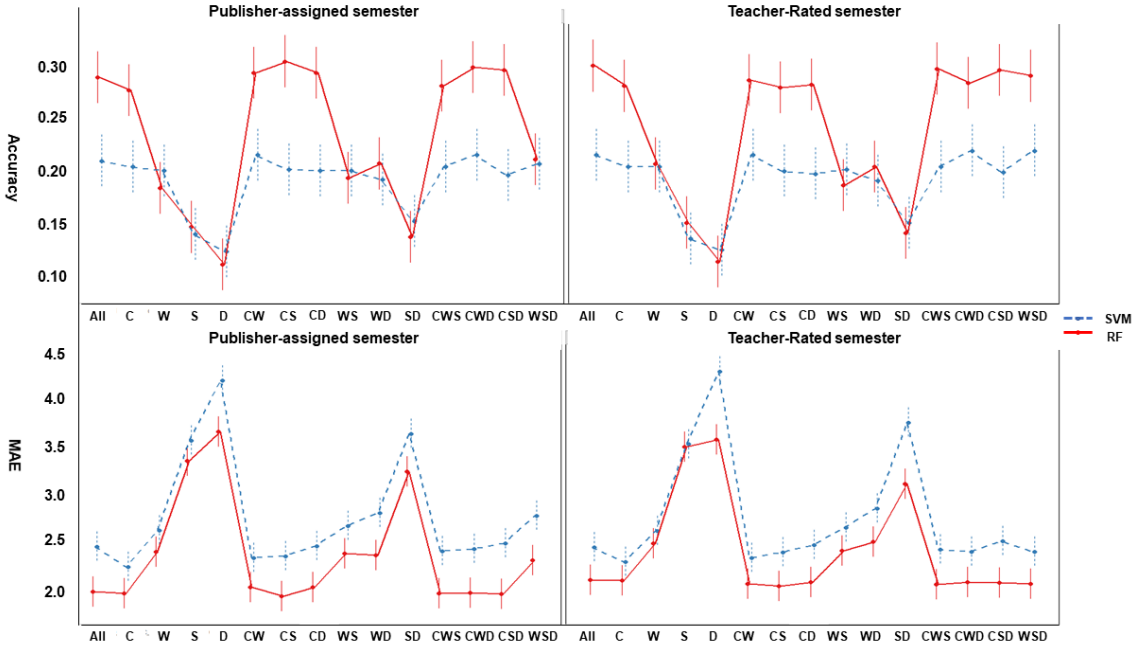
Примечание. SVM = метод опорных векторов; RF = случайный лес; C = Особенности символов; W = Особенности слов; S = Синтаксические особенности; D = Особенности дискурса; All = объекты на всех уровнях; Буквосочетания представляют собой комбинацию объектов на разных уровнях, например, C_W = объекты на уровне символов и слов

Анализ значимости признаков в наилучших моделях

Линейные смешанные модели (LMM) показали, что наилучшими моделями для предсказания читаемости текста являются модели случайного леса (Random Forest, RF), использующие либо все лингвистические признаки (модель all-RF), либо только признаки на уровне иероглифов (character-RF). Анализ значимости признаков показал, что важнейшие переменные для обоих показателей читаемости — как назначенных издателями (Y1), так и оценённых учителями (Y2) — были схожими. Это указывает на то, что признаки на уровне иероглифов превосходят по значимости признаки на синтаксическом и дискурсивном уровнях. Наиболее важным оказалось наличие психолингвистической информации об иероглифах. В частности, ключевую роль во всех оптимальн ых моделях играли семестр и класс, в котором тот или иной иероглиф впервые появляется в

Рисунок 1

Результаты точности прогнозирования и средней абсолютной ошибки (MAE) в моделях читаемости с использованием различных лингвистических характеристик



Пометка. SVM = метод опорных векторов; RF = случайный лес; C = Символ; W = Слово; S = Синтаксис; D = Дискурс; All = Character_Word_Syntax_Дискурс; Буквосочетания представляют собой комбинацию признаков на разных уровнях, например, C_W = признаки на уровне символов и слов.

учебнике (в расчёте на общее число вхождений или уникальные иероглифы). В моделях character-RF для Y1 особенно значимыми оказались время реакции и частота ошибок при назывании иероглифов взрослыми носителями китайского языка (по данным Liu и др., 2007). Возраст усвоения (age of acquisition) оказался важным в трёх из четырёх моделей, оцениваемых по точности (accuracy), но не был значим в модели all-RF для Y2. Также в топ-10 признаков всех моделей (кроме двух моделей character-RF для Y2) вошли: наличие фонетических подсказок в иероглифе, лёгкость описания его значения, прозрачность семантического радикала. Кроме того, во всех оптимальных моделях отмечалась высокая значимость частотности иероглифов. Как суммарная, так и усреднённая частота использования иероглифов входила в число ведущих признаков. Однако в разных моделях использовались данные частот из разных корпусов. Отношение количества уникальных иероглифов (type) к общему числу вхождений (token) было отобрано только в модели character-RF для Y2 при оценке MAE (средней абсолютной ошибки). В отличие от моделей для Y1, в трёх из четырёх моделей для Y2 оказалось важным структурное усложнение иероглифа — то есть количество черт в его написании. Также были выявлены значимые признаки на уровне слов. В модели all-RF важную роль сыграли: суммарная частота одноиероглифных слов, количество различных типов слов, включая одноиероглифные, многоиероглифные и редкие слова. Два признака, связанные с синтаксическими категориями частей речи — абсолютное количество наречий и числительных — были отобраны только в модели all-RF для Y1 при расчёте MAE. Из признаков на дискурсивном уровне только один оказался значимым: в модели all-RF для Y2 при расчёте точности (ACC) была выделена доля всех возможных пар абзацев, содержащих одинаковые содержательные слова — как показатель референтной связности.

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Исследования, проведенные с использованием методов машинного обучения, продемонстрировали важность лингвистических особенностей на различных уровнях, таких как уровень слов, синтаксис и дискурс, для удобочитаемости текста (Фитцджеральд и др., 2015). В китайском языке исследования в Материковый Китай и Тайвань последовательно обнаруживали, что использованиефункций с нескольких уровней превосходит результаты использования функций с одного уровня (например, Liu etal., 2024; Sung et al., 2015; Wu et al., 2020). Это исследование было одним из первых , в котором изучалась читабельность текста в Гонконге. В нем было извлечено 274лингвистические характеристики из 723 оцифрованных учебников китайского языка и изобразительного искусства ,широко используемых в начальных школах Гонконга,представляют уровни символов, слов, синтаксиса и дискурса. Для изучения прогностических возможностей этих функций при оценке удобочитаемости текста были использованы два алгоритма машинного обучения, а именно SVM и RF. Настоящее исследование расширилопредыдущие иссле-

дования на китайском языке (например, Liu et al., 2024; S ung et al., 2015; Wu et al., 2020), сосредоточив внимание на более точной семестровой шкале удобочитаемости текста и введя субъективный индекс:семестровый уровень, присвоенный преподавателем, а также семестровый уровень, присвоенный издателем. Тем временем были определены важные лингвистические особенности для прогнозирования удобочитаемости текста в контексте Гонконга. Текущие результаты показали , что модели с функциями на уровне отдельных символов имодели с многоуровневыми функциями, включающими функции на уровне символов, показали те же результаты, что и модели со всеми 274 функциями. Результаты демонстрируют центральную роль характеристик символов в прогнозировании удобочитаемости текста на китайском языке. Результаты оценки важности признаков показали сходство между мнения издателей и преподавателей. Результаты, полученные с обеих точек зрения, показали, что решающее значение имели характеристики на уровне персонажа, то есть семестр и оценка, в течение которых персонаж впервые появляется в учебнике. Между тем, результаты , полученные с этих двух точек зрения, отличались. Модели , в которых семестры оценивалисьпреподавателями, подчеркивали важность количества штрихов, в то время как модели, в которых семестры оценивались издателями, подчеркивали влияние результатов исследований, то есть времени реакции взрослых и частоты ошибок при решении лексических задач (Liu et al., 2007).

Центральная роль признаков иероглифов в прогнозировании читаемости текстов на китайском языке

В соответствии с предыдущими исследованиями, проведёнными в материковом Китае (например, Wu и соавт., 2020) и Тайване (например, Sung и соавт., 2015), текущие результаты показывают, что лексические признаки (т.е. на уровне иероглифов и слов) оказываются более значимыми по сравнению с синтаксическими и дискурсивными признаками при оценке читаемости текста. В частности, модели, основанные как на уровне семестра, назначенном издателем, так и на уровне семестра, оценённом учителями, оказались похожими и показали наилучшие результаты по точности и средней абсолютной ошибке (MAE) при использовании только признаков на уровне иероглифов или всех 274 признаков. Модели случайного леса (RF) дополнительно подтвердили, что включение признаков уровня иероглифов повышает точность предсказания и снижает ошибку по сравнению с моделями без таких признаков. Эти сходные результаты свидетельствуют о том, что лексические признаки имеют более сильное влияние на читаемость текста, чем синтаксические и дискурсивные, во всех китайских сообществах, независимо от варианта письменного и устного языка. Хотя было выявлено, что признаки на уровне иероглифов и слов обладают большей предсказательной силой, это не означает, что признаки синтаксиса и дискурса не влияют на читаемость. Предыдущие исследования подчёркивают важность синтаксических и дискурсивных навыков для понимания прочитанного (Chik и соавт., 2012). Например, исследование с участием учеников четвёртых классов Гонконга (Yeung и соавт., 2013) показало, что после учёта навыков чтения слов, синтаксические (знание порядка слов, морфосинтаксические знания) и дискурсивные навыки (знание порядка предложений) вносят уникальный вклад в понимание текста. Следовательно, несмотря на значительную роль лексических признаков, синтаксические и дискурсивные признаки также важны.

С другой стороны, текущие результаты расходятся с данными предыдущих исследований в Тайване (Chen и соавт., 2013; Sung и соавт., 2015), где модели с признаками только одного уровня не показывали такой же высокой эффективности, как модели с признаками нескольких уровней. Это различие может объясняться особенностями построения моделей в настоящем исследовании: были использованы два алгоритма машинного обучения (RF и SVM), а также более детальная шкала оценки — по семестрам, а не по классам, как в прежних работах. Кроме того, в настоящем исследовании применялось больше лингвистических признаков (274), и признаки на уровне иероглифов и слов были разделены, чего не делали в предыдущих исследованиях. В отличие от нашего исследования, где признаки на уровне иероглифов оказались более информативными, исследование в материковом Китае (Wu и соавт., 2020), где используется упрощённое письмо, показало преимущество признаков на уровне слов. Высокая эффективность признаков иероглифов в нашем исследовании может быть связана с использованием традиционного китайского письма в Гонконге (McBride-Chang и соавт., 2005). Традиционные иероглифы, имеющие идеографическое происхождение, сохраняют более тесную связь между формой и значением. Процесс упрощения в материковом Китае ослабляет эту связь, что может затруднять распознавание и чтение упрощённых иероглифов, особенно для начинающих читателей. Например, упрощенный иероглиф 爱 (любовь) был разработан путем удаления элемента, связанного со значением всего иероглифа, т.е. 心/sam1/ (сердце), из его традиционного аналога 愛/oi3/. Исследования показали , что дети, изучающие упрощенные символы, лучше справляются с заданиями на развитие зрительных навыков по сравнению с теми, кто изучает традиционные символы (например, McBride-Chang et al., 2005). Это говорит о том, что некоторые признаки на уровне иероглифов делают традиционные иероглифы относительно более лёгкими для распознавания и чтения, что в свою очередь влияет на понимание текста. Таким образом, различия между нашим исследованием и исследованиями в Тайване и материковом Китае могут быть обусловлены языковыми и письменными особенностями регионов, что подчёркивает необходимость создания специальных формул оценки читаемости для разных китайских сообществ.

Сходства и различия в роли признаков с точки зрения издателей и учителей

Настоящее исследование развивает предыдущие работы (например, Liu и соавт., 2024; Sung и соавт., 2015; Wu и соавт., 2020), включив в анализ как уровень семестра, назначенный издателем, так и уровень семестра, оценённый учителями, в качестве показателей читаемости текста. В исследовании были выявлены согласованные результаты по важности признаков как с точки зрения издателей, так и учителей. В частности, семестр и класс, в котором впервые вводится иероглиф в учебнике, оказывают значительное влияние на читаемость текста. Это влияние сохраняется значимым даже при учёте всех признаков, охватывающих четыре лингвистических уровня. Кроме того, возраст усвоения иероглифа, тесно связанный с семестром и классом его введения, также был обнаружен как значимый фактор, влияющий на читаемость. Эти признаки отражают последовательность изучения иероглифов, которая не является случайной. В Гонконге издатели обязаны при подготовке учебников опираться на Лексические списки для изучающих китайский язык в Гонконге (Education Bureau, 2007).

Согласно этому документу, символы, которые появляются на ранних этапах обучения, обычно имеют меньшую визуальную сложность и более высокую частотность, например, 一/jat1/ (один), 我/ngo5/ (я) и 你/nei5/ (ты), чем те, которые обычно изучаются позже, например, 勢/sai3/ (сила), 滲/sam3/ (проникновение) и 癒/jyu6/ (исцеление). Это говорит о том, что персонажи, которым учат на начальных этапах, должны быть проще, чем те , которые будут представлены позже. Более того, персонажи, которым учат раньше, могут быть более понятными, что позволит детям лучше понимать их с помощью контекстуального чтения (Brent & Siskind, 2001). Следовательно, дети могут освоить символы, с которыми они познакомились в раннем возрасте, что улучшит читаемость текста.

В соответствии с Лю и соавторами (2024), анализ значимости признаков выявил важность частоты встречаемости символов и слов в удобочитаемости текста. Это согласуется с предыдущими исследованиями, показывающими сильный частотный эффект, при котором высокочастотные слова читаются более точно и быстрее на нескольких языках (например, Cai & Brysbaert, 2010). Было высказано предположение, что более высокие частоты могут облегчить понимание символов и слов. Между тем, между нынешними результатами было несколько различий в отношении важности функции для модели семестров, назначенные издателями, и модели семестров, оцененные преподавателями. В частности, характеристики количествоштрихов, которые коррелируют с визуальной сложностью символов, занимают видное место в топ-10 характеристикоптимальных моделей для семестров, оцениваемых преподавателями, но не в моделях для семестров, назначаемых издателями. Наш анализ значимости признаков выявил меньшее количество признаков, связанных со сложностью, по сравнению с частотностью, что свидетельствует об относительно незначительном влиянии сложности на читаемость. В соответствии с этим, исследование китайских детей (Su & Samuels, 2010) было обнаружено уменьшение влияния визуальной сложности на обработку текста по мере развития навыков чтения у детей. С другой стороны, особенности, связанные со временем ответа взрослых и частотой ошибок при решении лексических задач (Liu et al., 2007), были обнаружены только в ходе анализа семестров, назначенных издательством. Это расхождение может быть объяснено относительной удобочитаемостью для учителей при непосредственном восприятии информации о времени реакции взрослых и частоте ошибок при решении лексических задач по сравнению с количеством штрихов.Следовательно, в то время как издатели могут полагаться на различные источники информации при распределении семестров, преподаватели, скорее всего, предпочтутиспользовать индексы, которые могут быть непосредственно получены из самих текстов , чтобы оценить уровень читаемости.

Ограничения и направления для будущих исследований

Поскольку данное исследование является одним из первых, посвящённых читаемости текстов в Гонконге, в используемом корпусе рассматривались только учебники трёх издательств. В будущих работах можно расширить разнообразие текстов, включив, например, художественные книги. Несмотря на то что мы привлекли опытных учителей для оценки текстов с учётом среднего уровня читателя определённого класса, всё же остаётся сложной задачей отразить реальную читаемость текстов для самих детей. В будущих исследованиях целесообразно учитывать оценки, данные непосредственно детьми, а также их результаты по заданиям на понимание прочитанного, которые напрямую связаны с показателями читаемости текста (Mesmer & Hiebert, 2015).

Кроме того, в будущем могут быть использованы дополнительные алгоритмы машинного обучения, подходящие для задач классификации, такие как метод k-ближайших соседей или алгоритм построения дерева решений (Rodriguez-Galiano и др., 2015). На первоначальных этапах точность моделей была невысокой, однако она улучшалась, когда уровень читаемости определялся по классам (с 1 по 6 класс). В частности, алгоритмы SVM и Random Forest показали схожие результаты. Обе модели — как с использованием только признаков на уровне иероглифа (SVM: средняя точность = 66,08%, SD = 0,04; RF: средняя точность = 70,94%, SD = 0,04), так и со всеми признаками (SVM: средняя точность = 65,57%, SD = 0,04; RF: средняя точность = 68,34%, SD = 0,04) — показали хорошую производительность и превзошли модели, не включавшие признаки на уровне иероглифа. Таким образом, полученные показатели точности сопоставимы с результатами предыдущих исследований, проведённых в материковом Китае (например, Wu и др., 2020) и на Тайване (например, Sung и др., 2015), что вносит вклад в общее

понимание читаемости текстов среди китайскоязычных сообществ. Однако это также подчёркивает необходимость дальнейшего изучения моделей с более детализированными шкалами, которые смогут обеспечивать более высокую точность предсказания уровня читаемости. Будущие исследования должны быть направлены именно на разработку таких моделей.

ЗАКЛЮЧЕНИЕ

Основной целью данного исследования было изучение прогностической силы лингвистических признаков на уровнях иероглифа, слова, синтаксиса и дискурса при отнесении текстов к определённым семестрам начальной школы. С помощью надёжных методов машинного обучения было показано, что лингвистические признаки, особенно на уровне иероглифа, обладают высокой предсказательной способностью. Кроме того, в рамках дополнительной цели исследования были проанализированы две оптимальные модели Random Forest — одна на основе всех признаков, другая только на основе признаков уровня иероглифа. Обе модели показали высокую точность и низкую среднюю абсолютную ошибку (MAE) при прогнозировании семестровой сложности текста. Анализ важности признаков выявил, что особенно значимыми являются последовательность изучения иероглифов, частотность их употребления, а также частота встречаемости слов. Эти результаты напрямую отвечают на исследовательские вопросы, определяя ключевые лингвистические характеристики, влияющие на оценку читаемости текста как со стороны издателей, так и со стороны учителей. С практической точки зрения полученные данные представляют ценность для образовательной практики. Учителя могут сосредоточить внимание на признаках лексического уровня, особенно при обучении новым иероглифам. Кроме того, в будущих исследованиях можно разработать автоматический анализатор читаемости текстов, ориентированный на признаки уровня иероглифа и основанный на двух оптимальных моделях RF, выявленных в настоящем исследовании. Такой инструмент может упростить процесс присвоения текстам семестровой сложности и помочь определять уровень читаемости материалов из других источников, например, художественных книг. В результате дети, родители и учителя смогут легче подбирать как формальные, так и неформальные тексты, соответствующие уровню чтения ребёнка.

БЛАГОДАРНОСТИ

Эта работа была частично поддержана фондом ResearchSeed Департамента специального образования и консультирования Образовательного университета Гонконга (регистрационный номер 04670) доктору Дью Лю. Эта работа также была частично поддержана присуждениемстипендии доктору Чжэнъе Сюй из Исследовательского центра. Совет по грантам Специального административного района Гонконг, Китай (проект № EdUHK PDFS2122-8H09).

ЗАЯВЛЕНИЯ О ДОСТУПНОСТИ ДАННЫХ

Данные, подтверждающие выводы этого исследования, доступны на https://osf.io/h9ew4/

КОНФЛИКТ ИНТЕРЕСОВ

Авторы заявляют об отсутствии конфликта интересов.

ВКЛАД АВТОРОВ

Чжэнъе Сюй: концептуализация; обработка данных; формальный анализ; привлечение финансирования; методология; администрирование проекта; программное обеспечение; авторский надзор; визуализация; написание оригинального проекта; рецензирование и редактирование.

Исюнь Ли: сбор данных; исследование; методология; ресурсы; написание – первоначальный вариант.

Дью Лю: концептуализация; сбор данных; привлечение финансирования; исследование; методология; администрирование проекта; программное обеспечение; написание – первоначальный вариант; рецензирование и редактирование.

ЛИТЕРАТУРА

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390-412. https://doi.org/10.1016/j.jml.2007.12.005

Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition, 81*(2), B33-B44. https://doi.org/10.1016/S0010-0277(01)00122-6

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles [Data set]. *PloS One, 5*(6), Article e10729. https://doi.org/10.1371/journal.pone.0010729

Chen, Y., Chen, Y., & Cheng, Y. (2013). Assessing Chinese readability using term frequency and lexical chain. *Journal of Computational Linguistics & Chinese Language Processing, 8*(2), 1-18.

Chen, Y., Tsai, Y., & Chen, Y. (2011). Chinese readability assessment using TF-IDF and SVM. 2*011 International Conference on Machine Learning and Cybernetics* (pp. 705-710). IEEE. https://doi.org/10.1109/ICMLC.2011.6016783

Cheng, Y., Xu, D., & Dong, J. (2020). 基于语文教材语料库的文本阅读难度分级关键因素分析与易读性公式研究 [A study on the analysis of key factors of text reading difficulty grading and the readability formula based on a corpus of language teaching materials]. 语言文字应用*, 1*, 132-143. https://doi.org/10.16499/j.cnki.1003-5397.2020.01.014

Chik, P. P., Ho, C. S., Yeung, P., Chan, D. W., Chung, K. K., Luan, H., Lo, L., & Lau, W. S. (2012). Syntactic skills in sentence reading comprehension among Chinese elementary school children. *Reading and Writing, 25*, 679-699. https://doi.org/10.1007/s11145-010-9293-4

Crossley, S. A., Skalicky, S., & Dascalu, M. (2019). Moving beyond classic readability formulas: New methods and new models. *Journal of Research in Reading, 42*(3-4), 541-561. https://doi.org/10.1111/1467-9817.12283

Crossley, S., Heintz, A., Choi, J. S., Batchelor, J., Karimi, M., & Malatinszky, A. (2023). A large-scaled corpus for assessing text readability. *Behavior Research Methods, 55*(2), 491-507. https://doi.org/10.3758/s13428-022-01802-x

Da, J. (2004). A corpus-based study of character and bigram frequencies in Chinese e-texts and its implications for Chinese language instruction [Data set]. *Proceedings of the Fourth International Conference on New Technologies in Teaching and Learning Chinese,* 501-511.

Education Bureau. (2007). *Lexical Lists for Chinese Learning in Hong Kong*.

Fitzgerald, J., Elmore, J., Koons, H., Hiebert, E. H., Bowen, K., Sanford-Moore, E. E., & Stenner, A. J. (2015). Important text characteristics for early-grades text complexity. *Journal of Educational Psychology, 107*(1), 4–29. https://doi.org/10.1037/a0037289

Givón, T. (1995). Coherence in text vs. coherence in mind. *Coherence in Spontaneous Text, 1995*, 59-116.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-metrix: Providing multilevel analyses of text characteristics. *Educational Researcher, 40*(5), 223-234. https://doi.org/10.3102/0013189X11413260

Huang, C. (2006). Automatic acquisition of linguistic knowledge: From sinica corpus to gigaword corpus [Conference presentation] [Data set]. *The 13th National Institute of Japanese Language International Symposium Language Corpora: Their Compilation and Application*. Tokyo.

Jin, G. J., Xiao, H., Fu, L., & Zhang, Y. F. (2005). 现代汉语语料库建设及深加工 [Construction and further processing of Chinese National Corpus] [Data set]. 语言文字应用*, 2*, 111–120. https://doi.org/10.16499/j.cnki.1003-5397.2005.02.017

Jing, X. (1995). 中文国文教材的适读性研究： 适读年级值的推估 [A study on the readability of Chinese national language teaching materials: Estimation of readability values of grade levels]. 教育研究资讯*, 5*, 113-127.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Flesch-kincaid grade level*. United States Navy.

Korobov, M., & Lopuhin, K. (2019). *Permutation importance.*

Liu, M., Li, Y., Su, Y., & Li, H. (2024). Text complexity of Chinese elementary school textbooks: Analysis of text linguistic features using machine learning algorithms. *Scientific Studies of Reading, 28*(3), 235-255. https://doi.org/10.1080/10888438.2023.2244620

Liu, M., Li, Y., Wang, X., Gan, L., & Li, H. (2021). 分级阅读初探: 基于小学教材的汉语可读性公式研究 [Leveled reading for primary students: Construction and evaluation of Chinese reeadability formulas based on textbooks]. 语言文字应用*, 2*, 116-126. https://doi.org/10.16499/j.cnki.1003-5397.2021.02.010

Liu, Q., Zhang, H. P., Yu, H. K., & Cheng, X. Q. (2004). 基于层叠隐马模型的汉语词法分析 [Chinese lexical analysis using cascaded hidden Markov model]. 计算机研究与发展*, 41*(8), 1421–1429.

Liu, Y., Shu, H., & Li, P. (2007). Word naming and psycholinguistic norms: Chinese [Data set]. *Behavior Research Methods, 39*(2), 192–198. https://doi.org/10.3758/BF03193147

Ma, W., & Chen, K. (2005). Design of CKIP Chinese word segmentation system. *Chinese and Oriental Languages Information Processing Society, 14*(3), 235-249.

McBride-Chang, C., Chow, B. W., Zhong, Y., Burgess, S., & Hayward, W. G. (2005). Chinese character acquisition and visual skills in two Chinese scripts. *Reading and Writing, 18*, 99-128. https://doi.org/10.1007/s11145-004-7343-5

Mesmer, H. A. E. (2005). Text decodability and the first-grade reader. *Reading & Writing Quarterly, 21*(1), 61-86. https://doi.org/10.1080/10573560590523667

Mesmer, H. A., & Hiebert, E. H. (2015). Third graders' reading proficiency reading texts varying in complexity and length: Responses of students in an urban, high-needs school. *Journal of Literacy Research, 47*(4), 473-504. https://doi.org/10.1177/1086296X16631923

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel. M., Prettenhofer. P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research, 12*, 2825-2830.

Pinney, C., Kennington, C., Pera, M. S., Wright, K. L., & Fails, J. A. (2024). Incorporating word-level phonemic decoding into readability assessment. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, Italia (pp.* 8998-9009). ELRA and ICCL.

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., & Chica-Rivas, M. J. O. G. R. (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews, 71*, 804-818. https://doi.org/10.1016/j.oregeorev.2015.01.001

Solnyshkina, M., Zamaletdinov, R., Gorodetskaya, L., & Gabitov, A. (2017). Evaluating text complexity and Flesch-Kincaid grade level. *Journal of Social Studies Education Research, 8*(3), 238-248.

Stutz, F., Schaffner, E., & Schiefele, U. (2016). Relations among reading motivation, reading amount, and reading comprehension in the early elementary grades. *Learning and Individual Differences, 45*, 101-113. https://doi.org/10.1016/j.lindif.2015.11.022

Su, I., Yum, Y. N., & Lau, D. K. (2023). Hong Kong Chinese character psycholinguistic norms: Ratings of 4376 single Chinese characters on semantic radical transparency, age-of-acquisition, familiarity, imageability, and concreteness [Data set]. *Behavior Research Methods, 55*(6), 2989-3008. https://doi.org/10.3758/s13428-022-01928-y

Su, Y., & Samuels, S. J. (2010). Developmental changes in character-complexity and word-length effects when reading Chinese script. *Reading and Writing, 23*, 1085-1108. https://doi.org/10.1007/s11145-009-9197-3

Sung, Y., Chen, J., Cha, J., Tseng, H., Chang, T., & Chang, K. (2015). Constructing and validating readability models: The method of integrating multilevel linguistic features with machine learning. *Behavior Research Methods, 47*, 340-354. https://doi.org/10.3758/s13428-014-0459-x

Tsang, Y., Huang, J., Lui, M., Xue, M., Chan, Y. F., Wang, S., & Chen, H. (2018). MELD-SCH: A megastudy of lexical decision in simplified Chinese [Data set] . *Behavior Research Methods, 50*, 1763-1777. https://doi.org/10.3758/s13428-017-0944-0

Tse, C., Yap, M. J., Chan, Y., Sze, W. P., Shaoul, C., & Lin, D. (2017). The Chinese lexicon project: A megastudy of lexical decision performance for 25,000 traditional Chinese two-character compound words. *Behavior Research Methods, 49*, 1503-1519. https://doi.org/10.3758/s13428-016-0810-5

Wang, F., & Wu, F. (2020). Postnominal relative clauses in Chinese. *Linguistics, 58*(6), 1501-1542. https://doi.org/10.1515/ling-2020-0226

Wu S., Yu D., & Jiang X. (2020). 汉语文本可读性特征体系构建和效度验证 [Development of linguistic features system for Chinese text readability assessment and its validity verification]. 世界汉语教学*, 34*(1), 81-97.

Yeung, S. S., Siegel, L. S., & Chan, C. K. (2013). Effects of a phonological awareness program on English reading and spelling among Hong Kong Chinese ESL children. *Reading and Writing, 26*, 681-704. https://doi.org/10.1007/s11145-012-9383-6

Zhang, J., McBride-Chang, C., Wong, A. M., Tardif, T., Shu, H., & Zhang, Y. (2014). Longitudinal correlates of reading comprehension difficulties in Chinese children. *Reading and Writing, 27*, 481-501. https://doi.org/10.1007/s11145-013-9453-4