Сравнительное исследование русских текстов ООН

Марико Мохамед Ламин, аспирант,
Казанский (Приволжский) федеральный университет

Аннотация: наше исследование направлено на построение лексической типологии русских текстов ООН путем определения диапазона метрик набора параметров, а именно: уровень плоти Кинкейда, лексической плотности, лексического разнообразия (TTR) и частоты встречаемости юридических терминов. Мы использовали данные с сайта ООН (https://www.ohchr.org/EN/PublicationsResources/Pages/RecentPublications.aspx) и составили корпус из 20 текстов ООН общим объемом около 20 000 слов. Показатель уровень плоти Кинкейда, который рассчитывается на основе средней длины предложения, количества слов и количества слогов, оказался относительно высоким и колеблется между 11,75 и 16,49. Результаты показывают, что тексты являются одновременно лексически плотными и разнообразными: лексическая плотность варьируется от 67 до 74, а TTR - от 0,56 до 0,62. Таким образом, они демонстрируют высокую степень корреляции с ФКГЛ (FKGL). Типичный русский текст ООН содержит около 66 юридических терминов, что показывает, что 7 текстов из 8 содержат частотность юридических терминов, варьирующуюся от 66 до 73. Мы пришли к выводу, что диапазон метрик четырех параметров, выбранных выше, может служить предикторами в определении сложности русских текстов ООН, тем самым подтверждая точку зрения Ure [31]. Полученные результаты могут представлять интерес для преподавателей, исследователей сложности текста и переводоведения.

Ключевые слова: уровень плоти Кинкейда или сложность текста, лексическая плотность, лексическое разнообразие (TTR), частота встречаемости юридических терминов, русские тексты ООН

Для цитирования: Марико Мохамед Ламин Сравнительное исследование русских текстов ООН // Modern Humanities Success. 2023. № 8. С. 57 – 69.

Поступила в редакцию: 28 мая 2023 г.; Принята в доработанном виде: 20 июня 2023 г.; Одобрена для публикации: 31 июля 2023 г.

Введение

Во-первых, оценка сложности текста является важной задачей как в образовании, так и в переводоведении. Во-вторых, она может помочь преподавателям выбрать подходящие тексты для учащихся с учетом их когнитивных способностей и уровня владения языком. В-третьих, оценка сложности текста может облегчить измерение качества перевода, что, в свою очередь, поможет сравнить или сопоставить многоязычные тексты. Поскольку знания передаются в основном в письменной форме, чтение становится одним из самых важных инструментов в процессе обучения.

Целью данной работы является построение лексической типологии русских текстов ООН по четырем параметрам, а именно: уровень плоти Кинкейда (FKGL) или сложность текста, лексическая плотность, лексическое разнообразие (TTR) и частота встречаемости юридических терминов. Цели данного исследования две: во-первых, проверить валидность автоматизированного профилировщика русских академических текстов *Rulingva* (https://Rulingva.kpfu.ru/), остановившись на четырех выбранных параметрах; во-вторых, построить лексическую типологию русских текстов ООН и найти метрический диапазон вышеупомянутых параметров. С учетом этих целей мы представляем новый алгоритм классификации русских текстов на основе четырех выбранных метрик. Исследовательский вопрос, на который опиралось настоящее исследование, был следующим: Каковы лексические типы русских текстов ООН?

Обзор литературы

В данном разделе рассматривается литература, связанная со сложностью текста и некоторыми его параметрами, включая уровень плоти Кинкейда, лексическую плотность, лексическое разнообразие и частоту встречаемости юридических терминов.

Сложность текста

Biber [2] утверждает, что сложность текста признана многомерной конструкцией и количественно измеряется с разных точек зрения Castello и других [4, 23, 25, 8, 28] В их исследовании сложность текста рассматривается с системно-функциональной точки зрения, где сложность текста оценивается по двум параметрам: грамматической запутанности и лексической плотности. Считается, что эти два измерения взаимозависимы Halliday [16, 14] поскольку грамматика и лексика рассматриваются одновременно Hasan [18]. Стоит отметить, что грамматическая сложность исключена из данного исследования, то есть мы сосредоточились на лексической плотности для выявления сложности в русских текстах ООН. Ниже мы приводим подробную информацию о четырех параметрах, выбранных для исследования.

Уровень плоти Кинкейда

Исследования сложности начались с формул читабельности. Что касается английского языка, то существует не менее 200 показателей удобочи-

таемости текстов, и все они в основном измеряют два аспекта предложения: сложность слова, обычно через длину слова или количество слогов, и сложность предложения, обычно через длину предложения Dubay [6]. В данном конкретном исследовании мы попытались соотнести наш корпус (тексты) с уровнем класса. Одним из старейших и хорошо известных методов определения соответствующего уровня учебника является формула Флеша-Кинкейда Klare [22]. В основе этого метода лежит формула, которая рассчитывает среднее количество слогов в слове и среднюю длину всех предложений в тексте. Формула вычисляется следующим образом: FK= (.39 x ASW) – 15.59

FK означает Flesh-Kincaid Grade Level, ASL – средняя длина предложения (количество слов, деленное на количество предложений), а ASW – среднее количество слогов в слове (количество слогов, деленное на количество слов). Эта формула элегантна в своей простоте. Она берет поверхностные характеристики текста и выдает индекс сложности, выраженный в виде уровня оценки. Кроме того, она эффективна. Длина предложения коррелирует с переменными, которые влияют на усилия, необходимые для прочтения предложения, такими как синтаксическая сложность. Количество слогов дает показатель длины слов, который обратно коррелирует с частотой слов и влияет на трудность чтения Zipf [35]. Мы не ожидаем увидеть много слов из четырех слогов в сборнике рассказов для детского сада. И наоборот, мы не ожидаем увидеть цепочки из трех или четырех слов в тексте по биологии для старшеклассников.

Что касается современных подходов, то уровень оценки по Флешу-Кинкейду и связанные с ним меры, такие как Flesh Reading Ease, имеют хорошо известные недостатки, и по этой причине ведется активная работа по разработке более совершенных методов определения уровня оценки. Одним из преимуществ уровня Флеш-Кинкейда является то, что количество слогов и длина предложения легко измеряются и коррелируют с такими важными конструктами, как сложность слова и сложность предложения. Однако сегодня эта простота вычисления является менее актуальной, чем раньше, благодаря достижениям в области вычислительной лингвистики.

Лексическая плотность

Вводя понятие лексической плотности, Ure [31] различает слова с лексическими свойствами (содержательные слова) и без них (функциональные слова). Согласно Ure, элементы, не обладающие лексическими свойствами, т.е. функциональные слова, могут быть описаны "чисто в терминах грамматики" (с. 445), что означает, что такие слова или элементы обладают более грамматико-синтаксической функцией, чем лексические элементы. Лексическая плотность определяется общим количеством слов с лексическими свойствами (лексем). В результате получается процентное соотношение токенов слов с лексическими свойствами для каждого текста в корпусе. Уре приходит к выводу, что значительное большинство устных текстов имеют лексическую плотность менее 40%, в то время как значительное большинство письменных текстов имеют лексическую плотность 40% и выше. Одно замечание здесь заключается в том, что эти цифры должны сильно зависеть от языка - язык с более связанной морфологией, вероятно, покажет большее количество содержательных слов.

В данном исследовании *лексическая плотность* рассматривается как термин, наиболее часто используемый для описания количества слов содержания к общему количеству слов (лексем). Слова содержания важны для объяснения информации, поскольку лексические единицы необходимы для расчета лексической плотности текста; слова содержания состоят из существительных, глаголов, прилагательных и наречий Johansson [21], но есть такие слова, как предлог, союз, вспомогательные глаголы, модальные глаголы, местоимения и артикли, которые не классифицируются как лексические единицы. *Существительные* используются для обозначения лиц, мест, вещей, чувств или идей, например, Doctor, Canada, Pen, Fear и т.д. Существительные обычно отвечают на вопросы кто или что. *Глаголы* – это слова, которые показывают действие или поступок. *Прилагательные* – это слова, которые изменяют значение существительного Stern [30], то есть прилагательные – это слова, которые используются с существительным для описания или указания на человека, животное, место или вещь, которую называет существительное, или говорят о числе или количестве. Например, The *big red flashy* car. Согласно Stern [30], *наречие* является модификатором широкого диапазона – оно модифицирует все и вся, кроме существительных и местоимений. Johansson [21] утверждает, что наречия считаются лексическими единицами – это все наречия, образованные от прилагательных. Например: *ясно, печально, аккуратно* и т.д.

Было предложено несколько вариантов лексической плотности. Популярный "незначительный вариант" – подсчет плотности существительных, количество существительных, деленное на общее количество слов (лексем) в тексте. В более поздней статье Уре определяет лексическую плотность как "соотношение слов с лексическими значениями (члены открытых множеств) к словам с грамматическими значениями (элементы, представля-

ющие термины в закрытых множествах)". Поскольку все слова имеют грамматические значения, это отношение "часть: целое" Ure [32].

Ure [32] высказали мнение, что вопрос лексичности является жизненно важным при обсуждении понятия лексической плотности. Традиционно существительные, глаголы и прилагательные – это три класса слов, которые считаются обладающими лексическими свойствами. Под лексическими свойствами Ure [32] подразумевали слова содержания или слова открытого класса (из-за возможности легко включать новые члены класса – в то время как более грамматические части речи называются закрытыми классами, поскольку новые предлоги или местоимения редко входят в язык).

Концепция лексической плотности была разработана и уточнена Halliday [12]. Он подчеркивает важность разграничения между лексическими и грамматическими элементами. Элемент может состоять из более чем одного слова. Так, Halliday [12] считает *turn up* одним лексическим элементом, в то время как Ure [31] считает его одним лексическим элементом (*turn)* и одним грамматическим элементом (*up)*. Лексический элемент определяется Halliday как элемент, который "функционирует в лексических наборах, а не в грамматических системах: то есть вступает в открытые, а не закрытые контрасты" Halliday [12]. Лексический элемент является частью открытого множества, которое может быть противопоставлено ряду предметов в мире. Грамматический элемент, с другой стороны, входит в закрытую систему, согласно M.A.K. Halliday. Для грамматической системы характерно то, что входящие в нее классы (слов) имеют фиксированный набор элементов, в который невозможно добавить новые члены. С одной стороны, встречаются такие предлоги, как в (букв.: 'в'); на (букв.: 'на'); по (букв.: 'на, вдоль'); С (букв.: 'с'); за (букв.: 'для'); от (букв.: 'от'); при (букв.: 'при') и так далее. Когда вышеупомянутые предлоги используются в предложных фразах, идиоматических выражениях, составных предлогах, сложных предлогах, фиксированных сочетаниях (таких как: "в отношении" 'в связи', "в целях" 'в отношении', "со стороны" 'с одной стороны', "за счет" 'посредством'), а также парные предлоги ("из-за"), они имеют более конкретные, менее разнообразные значения, что снижает вероятность ошибок. С другой стороны, есть "вне" (дословно: "вне, за пределами"); "внутри" (дословно: "внутри, внутри"); "напротив" (перед); "около" (рядом); "после" (после) "прежде" (перед); "против" (против) "среди" (среди) и др. которые являются наречными предлогами, обозначающими наречное значение. Таким образом, на основании приведенных выше примеров можно утверждать, что предлоги могут выполнять как лексические, так и грамматические функции в зависимости от контекста, в котором они используются. Однако в данном исследовании основное внимание уделяется лексическим, а не грамматическим элементам.

По мнению M.A.K. Halliday, детский язык свидетельствует о существовании двух классов – лексического и грамматического. В начале своего языкового развития дети часто строят предложения, в которых отсутствуют все грамматические элементы. Кроме того, Halliday подчеркивает, что существует континуум от лексики к грамматике, и что есть – и всегда будут – промежуточные случаи. Например, он утверждал, что английские предлоги и некоторые классы наречий находятся на границе между лексическими и грамматическими элементами. Он приводит примеры таких наречий, как *always* и *perhaps*. При сравнении устной и письменной речи очень важно быть последовательным в проведении границы между "лексическими" наречиями, которые являются неграмматикализованными наречиями (включая все наречия, образованные от прилагательных), и "грамматическими" наречиями (наречия, входящие в закрытый класс предметов), но не так важно, где проводится эта граница.

Определение лексической плотности, данное Halliday, такого: "количество лексических единиц в пропорции к количеству употребляемых слов" Halliday [12]. Разница между определениями лексической плотности Halliday и Ure заключается в том, что Halliday считает некоторые наречия лексическими единицами. Таким образом, грамматические наречия включаются в элементы закрытого класса, а неграмматические наречия (включая все наречия, образованные от прилагательных) считаются лексическими элементами. В нашем исследовании лексическая плотность рассчитывалась путем деления количества лексических единиц на общее количество слов в каждом тексте. Eggins [7] указывает, что есть две основные лингвистические особенности, которые очень чувствительны к вариациям режима (*режим* относится к тому, как используется язык, является ли канал коммуникации устным или письменным и используется ли язык как способ действия или опровержения) степень грамматической сложности и лексическая плотность выбранного языка. Эти особенности отвечают, пожалуй, за самые разительные различия между устной и письменной речью. Halliday в своей книге "*Разговорный* и *письменный язык"* [12] также объясняет существенные различия между письменным и разговорным языком. Первое из них – это *плотность,* плотность, с которой представлена информация. По отношению друг к другу устный язык скуден (представленная часть

информации либо недостаточна, либо недостаточна), а письменный язык плотен (представленная информация обширна, другими словами, хорошо детализирована для понимания). Второе – это *замысловатость,* сложность, с которой организована идея в конкретном тексте. Разговорный язык более замысловатый, чем письменный. Кроме того, Halliday [17] утверждается, что письменный язык обычно является сложным (трудным), когда он имеет высокую лексическую плотность. Он хранит большое количество лексических единиц в каждой клаузе. С другой стороны, разговорный язык становится сложным (трудным), будучи грамматически запутанным. Стоит отметить, что под *сложностью* в нашем исследовании понимается лексическая сложность, а исследование сосредоточено на письменных текстах, а не на устных.

Halliday [14], отметил, что мы получаем понятие упаковки информации; текст с большим количеством слов содержания содержит больше информации, чем текст с большим количеством функциональных слов (предлогов, междометий, местоимений, союзов).

Лексическое разнообразие

Лексическое разнообразие измеряет, насколько сильно варьируется словарный запас языкового образца или текста. Для того чтобы текст отличался высоким лексическим разнообразием, говорящий или пишущий должен использовать много разных слов, с небольшим количеством повторений уже использованных слов. Соотношение типов слов, т.е. традиционная мера лексического разнообразия, представляет собой отношение количества различных слов (типов) к общему количеству слов (токенов), так называемое соотношение тип-токен, или TTR Malvern [24]. Проблема с показателем TTR заключается в том, что образцы текста, содержащие большое количество лексем, дают более низкие значения TTR и наоборот. Это объясняется тем, что количество словесных лексем может увеличиваться бесконечно, хотя то же самое верно и для типов слов. Часто для пишущего или говорящего важно повторно использовать несколько функциональных слов, чтобы составить одно новое лексическое слово. Это означает, что длинный текст в целом имеет более низкое значение TTR, чем короткий текст, что делает особенно сложным использование TTR в сравнении развития, например, между возрастными группами, где количество словесных лексем часто увеличивается с возрастом. Gayraud [9] сравнивает TTR и количество словесных лексем и указывает, что хотя количество словесных лексем значительно увеличивается с возрастом говорящего или пишущего, TTR падает. Таким образом, TTR можно использовать только при сравнении текстов одинаковой длины. Несмотря на это, TTR по-прежнему используется для сравнения текстовой продукции, например, между детскими текстами или между различными группами с языковыми нарушениями.

Лексическое разнообразие или лексическое богатство Daller [5] – это термины, которые относятся к статистическим показателям, измеряющим лексическое богатство текстов, а также могут использоваться для оценки общего прогресса учащихся. Лексическая насыщенность текста показывает, сколько различных слов используется в тексте, а лексическая плотность – долю лексических единиц, то есть существительных, глаголов, прилагательных и некоторых наречий в тексте Johansson [21]. Оба показателя применяются в компьютерном анализе корпусных данных. Как правило, тексты с меньшей плотностью легче воспринимаются, а устные тексты имеют более низкий уровень лексической плотности, чем письменные Ure and Halliday [31, 12]. Однако в Johansson [21] утверждается, что текст может иметь высокое лексическое разнообразие (т.е. содержать много различных типов слов), но низкую лексическую плотность, т.е. содержать много местоимений и вспомогательных слов, а не существительных и лексических глаголов или наоборот.

Частота встречаемости юридических терминов

Большое внимание также уделяется использованию юридических терминов, поскольку корпус основан на русских текстах ООН. Насколько нам известно, ни в одном из предыдущих исследований не обрабатывались русские тексты ООН, однако в Nation [26] представлен подробный обзор предыдущих исследований по объему словарного запаса и охвату текста. Например, они ссылаются на исследование Goulden [11], которое показало, что выпускник университета понимает около 20 000 "семейств слов". Учащиеся с гораздо меньшим объемом словарного запаса могут довольно успешно читать многие тексты. Например, Hirsh [19] обнаружили, что 2 000 наиболее распространенных семейств слов обеспечивают 90% охвата корпуса подростковых романов. В соответствии с приведенными выше объяснениями, мы связываем комплексность в русских текстах ООН как с уменьшением, так и с увеличением частотности юридических терминов. То есть, более высокая или более низкая частотность юридических терминов в тексте может влиять на понимание.

В заключение данного раздела мы утверждаем, что уровень плоти Кинкейда имеет некоторые ограничения, поскольку он охватывает только некоторые области, такие как длина предложения, количество слогов в предложении и т.д. Таким образом, мы использовали некоторые другие па-

раметры, а именно: лексическую плотность, лексическое разнообразие и частоту встречаемости юридических терминов, чтобы использовать их в качестве предикторов при определении сложности текста.

Материал и методы

Для данного исследования мы использовали вторичные данные с сайта ООН, создав корпус, состоящий примерно из 20 000 слов. Выборки ограничены 1 000 слов для создания двадцати (20) текстов, которые обозначены как русский текст один (RT1). Русский текст двадцать (RT20). Для описания мы выбрали четыре (4) параметра, которые являются следующими: уровень плоти Кинкейда; лексическая плотность; лексическое разнообразие (TTR) и частота встречаемости юридических терминов.

Насколько нам известно, ни в одном из предыдущих исследований не рассматривались русские тексты ООН. Поэтому мы ограничили наши выборки 1000 словами, основываясь на аргументе Biber [1] о том, что корпус должен быть достаточно большим, чтобы адекватно представлять изучаемые особенности. Например, при изучении лексики размер корпуса является еще более важным фактором. Распределение словарного запаса, описывающее количество слов, не является линейным. Это связано с тем, что слова имеют тенденцию повторяться в корпусе, и чем больше корпус, тем больше повторяющихся слов он содержит. Например, мы можем найти 500 типов слов в тексте объемом 1 000 слов, но очень маловероятно, что мы найдем 5 000 типов слов в корпусе объемом 10 000 слов, и невозможно найти 500 000 типов слов в корпусе объемом 1 миллион слов Следовательно, словарный запас имеет нелинейное распределение: по мере увеличения объема корпуса мы обнаруживаем лишь пропорционально небольшое увеличение количества новых типов слов.

Категория текста (также известная как жанр) относится к типу текста, например, является ли текст преимущественно повествовательным (например, романы, народные сказки), излагающим (например, учебники, журнальные статьи), убеждающим (например, редакционные статьи, проповеди) или описательным Biber, Zucker [3, 20].

В соответствии с вышеприведенными объяснениями, мы поддерживаем мнение Biber [1], который заключает, что образцы текста в 1000 слов являются репрезентативными для исследуемых категорий текста.

В результате трудно сравнить распределение словарного запаса в субкорпорах разного размера.

Инструмент

Rulingva (https://Rulingva.kpfu.ru/) – это автоматизированный инструмент, разработанный в Казанском федеральном университете (Российская Федерация); веб-инструмент, предназначенный для анализа русских учебных, профессиональных и промышленных текстов. Таким образом, *Rulingva* является инструментом, используемым в данном исследовании для расчета следующих параметров (лексическая плотность; лексическое разнообразие; уровень градации Флеша-Кинкейда; количество юридических терминов) русских текстов ООН.

Дизайн исследования

На первом этапе мы собрали данные с сайта ООН(https://www.ohchr.org/EN/PublicationsResources/Pages/RecentPublications.aspx), создав корпус русских текстов ООН. Наш корпус состоял примерно из 20 000 слов. Затем мы преобразовали тексты в текстовые файлы, удалив все необычные символы и изображения, чтобы их можно было использовать для нашего исследования. Biber [1] утверждается, что образцы текстов в 1000 слов являются репрезентативными для исследуемых категорий текстов. Поэтому мы отобрали из созданного нами корпуса образцы текстов объемом 1000 слов, чтобы получить 20 текстов.

На втором этапе мы вычислили выбранные тексты с помощью *Rulingva*, чтобы определить метрики выбранных параметров, т.е. Flesh-Kincaid Grade Level, лексическую плотность, лексическое разнообразие и частоту встречаемости юридических терминов.

Интерпретация данных

В таблице ниже представлены статистические данные двадцати (20) русских текстов ООН. Мы рассчитали лексическую плотность, разделив количество лексических единиц (существительных, глаголов, прилагательных и наречий) на общее количество слов, умноженное на 100. Например, RT1 состоит из существительных: 408; наречий: 40; прилагательных: 166; глаголов: 109, что в сумме составляет 723 ÷ (1013 = общее количество слов)×100=71,37%. Таким образом, 71,37% составляет лексическая плотность RT1. За ним следует Flesh Kincaid Grade level, который является показателем сложности текста; более подробные объяснения по этому поводу приведены выше. В данном исследовании большинство текстов варьируется между 11 и 14. Уровень FKGL немногих текстов, т.е. RT14; RT5; RT6; RT3 выше 14. Например, RT14 составляет 16,49, что говорит о том, что для понимания такого текста учащемуся требуется примерно 16 лет формального образования или больше. Мы также подсчитали количе-

ство юридических терминов, поскольку корпус состоит из юридических документов, а количество юридических терминов влияет на сложность.

Существуют различия в количестве лексической плотности каждого текста. В соответствии с утверждением Ure [31], если количество лексической плотности превышает 40%, это означает, что текст относится к категории письменного языка. Наше исследование показало, что все двадцать текстов превышают 40% по количеству лексической плотности, то есть варьируются от 68% до 74% (табл. 1), поэтому все двадцать текстов соответствуют характеристике письменного языка.

Выбор четырех (4) параметров обосновывается двумя причинами. Первая попытка заключается в построении лексической типологии русских текстов ООН для определения их сложности. Если наша гипотеза подтвердится, то вторая попытка будет заключаться в предположении, что выбранные параметры, т.е. Flesh Kincaid Grade Level, лексическая плотность, лексическое разнообразие и частотность юридических терминов, могут служить предикторами в определении сложности русских текстов ООН.

Результаты лексической плотности и лексического разнообразия были разделены на несколько категорий на основе классификации, предложенной экспертами (31). Что касается ФКГЛ и частотности юридических терминов, то они были рассчитаны с помощью программы *RuLingva.* Классификации таковы:

Таблица 1

Статистические данные четырех параметров (Лексическая плотность, Лексическое разнообразие, Уровень плоти Кинкейда и Юридические термины)

Русские тексты ООН	Лексическая плотность	Лексическое разнообразие (TTR)	Flesh-KGL	Юридические термины
1	71.37%	0.57	13.63	75
2	71.20%	0.58	14.43	86
3	70.20%	0.59	16.42	50
4	69.29%	0.6	14.37	59
5	72.40%	0.53	16.01	101
6	71.11%	0.54	16.07	101
7	69%	0.54	14.91	66
8	69.70%	0.52	15.88	95
9	69.24%	0.56	13.66	73
10	68.49%	0.57	12.32	109
11	69.42%	0.53	14.45	58
12	68.66%	0.51	14.21	67
13	71.05%	0.55	13.65	66
14	67.12%	0.56	16.49	29
15	74.65%	0.56	12.75	169
16	69.61	0.6	11.75	69
17	71.02%	0.62	15.63	66
18	70.53%	0.61	13.89	85
19	71.14%	0.55	14.24	69
20	68.48%	0.52	13.74	100

Примечание: *Диапазон лексической плотности*
41-50%: не плотный; 51-60% : менее плотный; 61-70% : плотный; >70% : очень плотный

Наш вывод согласуется с результатами других исследований, проведенных в области сложности текста. Мы приняли теорию Ure [31], согласно которой текст не является плотным, если он колеблется между 41% и 50%; менее плотным, если он колеблется между 51% и 60%; плотным, если он колеблется от 61% до 70%; очень плотным, если он выше 70%.

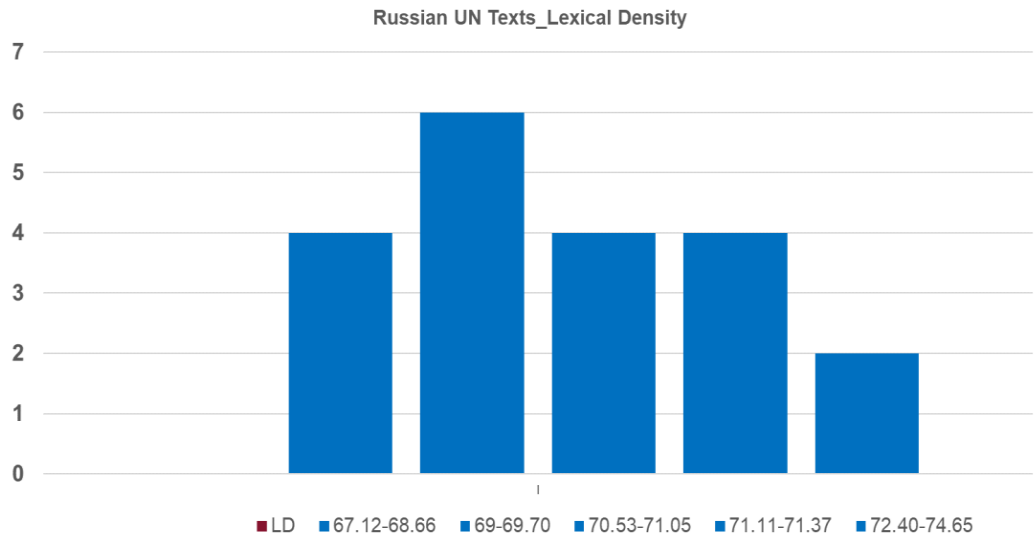

Рис. 1. Лексическая плотность русских текстов ООН

Рис. 1 показывает, что лексическое разнообразие вычисленных текстов колеблется между 67 и 71, что означает, что 6 текстов из 7 являются лексически плотными

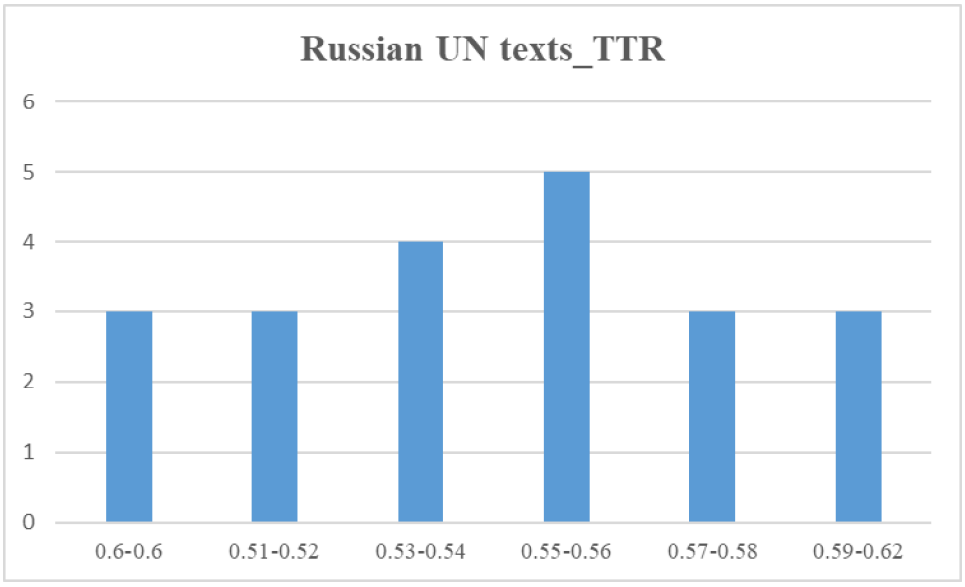

Рис. 2. Лексическое разнообразие (TTR)

Рис. 2 показывает лексическое разнообразие (TTR) вычисленных текстов, которое варьируется между 0,6 и 0,62. Он также показывает, что соотношение типов лексем в русских текстах ООН варьируется между 0,55 и 0,56.

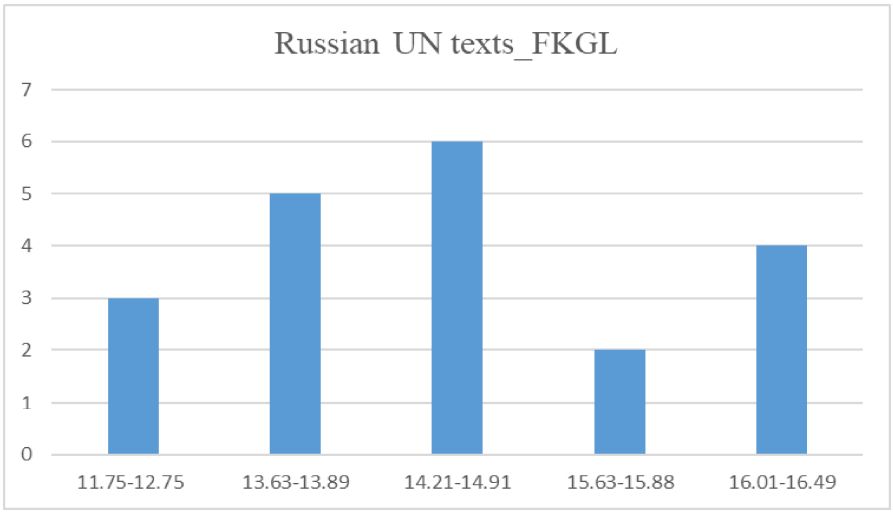

Рис. 3. Уровень плоти Кинкейда (FKGL)

На рис. 3 представлен Flesh Kincaid Grade Level, который колеблется между 11,75 и 16,49. FKGL русских текстов ООН колеблется между 14,21 и 14,91.

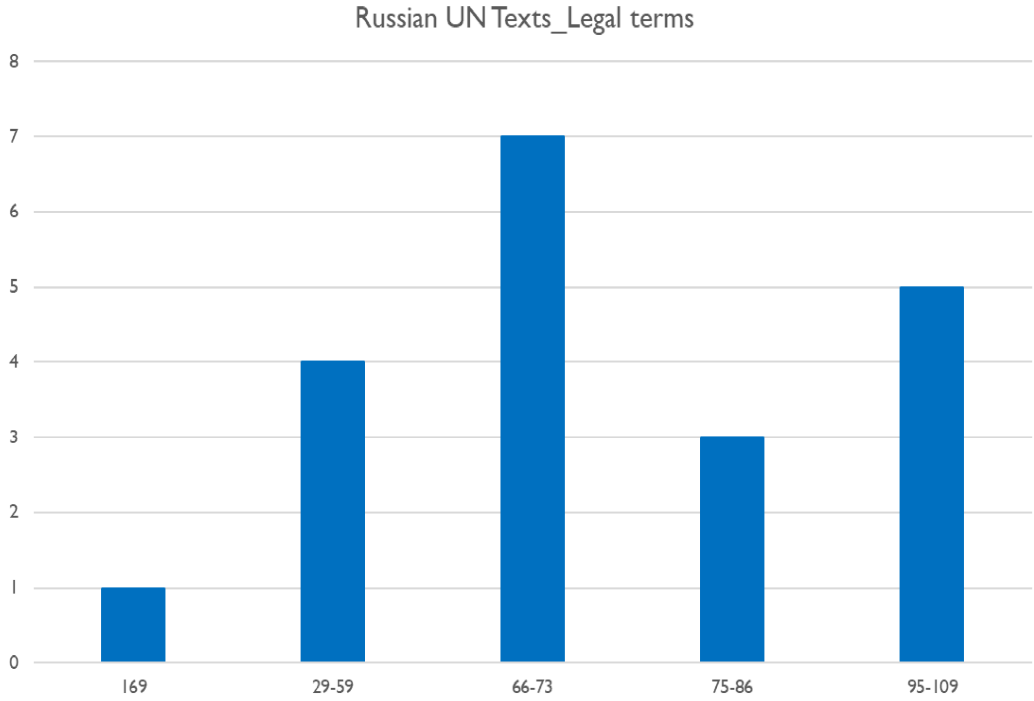

Рис. 4. Частота встречаемости юридических терминов

Рис. 4. Показывает, что типичный русский текст ООН содержит около 66 юридических терминов, что означает, что 7 текстов из 8 содержат юридические термины в диапазоне от 66 до 73.

Обсуждение

Наши результаты показали, что четыре (4) параметра, выбранные в данном исследовании, т.е. Flesh Kincaid Grade Level, лексическая плотность, лексическое разнообразие (TTR) и частота встречаемости юридических терминов, могут служить предикторами в определении сложности русских текстов ООН. Во-первых, лексическая плотность варьируется от 67% до 74%, что означает, что тексты являются лексически плотными, если обратиться к теории, принятой Ure [31], который указал, что текст является очень плотным, когда его лексическая плотность превышает 40%.

Во-вторых, мы рассчитали лексическое разнообразие (TTR). TTR – это измерение соотношения, когда количество различных типов делится на количество всех слов, или токенов. В результате расчета получается пропорция от 0 до 1, при этом более высокий показатель свидетельствует о большем разнообразии лексики в данной выборке Malvern [24]. Таким образом, TTR в нашем исследовании варьируется от 0,56 до 0,62, что считается относительно разнообразным, поскольку пропорция варьируется от 0 до 1. Теоретически, TTR

снижается из-за природы языка и повторения функциональной лексики, такой как предлоги и артикли. В Malvern [24] приводится теоретическая причина этого наблюдаемого явления: "Добавление дополнительного слова к языковой выборке всегда увеличивает количество лексем (N), но увеличивает количество типов (V), только если слово не использовалось ранее... Поэтому количество типов (V) в числителе увеличивается медленнее, чем количество лексем (N) в знаменателе, и TTR неизбежно падает" (стр. 22).

Как правило, тексты с меньшей лексической плотностью легче воспринимаются, а устные тексты имеют более низкий уровень лексической плотности, чем письменные Ure, Halliday [31, 12]. Таким образом, писатель мог использовать несколько синонимов с меньшим количеством повторений, что, вероятно, повлияло на сложность.

Далее следует FKGL, который варьируется от 11,75 до 16,49, а типичные тексты варьируются от 14,21 до 14,91. По мнению ряда авторов, различные наборы метрик для оценки сходства и несходства в сложности текста, такие как прилагательные на предложение, существительные на предложение, частота слов содержания и т.д., могут успешно ранжировать академические тексты для разных возрастных и классных уровней [29].

Основываясь на результатах, полученных в данном исследовании, мы можем повторить, что наши выводы подтверждают исследование, проведенное Solovyev [29], в котором авторы после анализа российских академических учебников по обществознанию утверждали, что минимальное значение уровень плоти Кинкейда составило 8,7, а максимальное – 28,09. Это означает, что значение уровень плоти Кинкейда более сложных текстов выше 16. Наконец, частота встречаемости юридических терминов колеблется между 66 и 73. Один текст из 8 содержит 169 юридических терминов. Это означает, что автор мог использовать повторяющиеся юридические термины для облегчения понимания, однако мы должны напомнить, что целевой аудиторией текстов ООН на русском языке являются политики, а учебники по русскому обществознанию предназначены для академических целей. Поэтому самый легкий для чтения русский текст ООН – 11,75 – не удивителен по сравнению с русскими академическими текстами по обществознанию "8,7" [29].

Исследование Yu [34] похоже на наше исследование в том, что авторы подчеркнули, как лексическая плотность помогает измерить сложность текста. Единственное различие между двумя исследованиями заключается в том, что авторы сосредоточились как на грамматической сложности, так и на лексической плотности, в то время как мы сделали ставку на лексическую плотность для определения сложности в русских текстах ООН. Так, они провели сравнительное исследование степени и вариативности сложности текста четырех английских переводов "*Сутры платформы",* которая была переведена как религиозный, так и литературный текст, причем стремление переводчика придать тексту литературный колорит проявилось во многих аспектах.

В соответствии с идеями, рассмотренными выше, и изучив работы различных ученых, мы можем сделать вывод, что ни в одном из предыдущих исследований не рассматривались русские тексты ООН. В результате мы сосредоточились на лексических элементах четырех параметров, выбранных выше, а не на функциональных элементах (см. обзор литературы). Исследование также доказало, что выбранные параметры могут служить предикторами при определении сложности в русских текстах ООН.

Заключение

В заключение следует отметить, что данное исследование было значимым, поскольку оно помогло нам прийти к выводу, что четыре выбранных параметра, т.е. уровень плоти Кинкейда, лексическая плотность, лексическое разнообразие (TTR) и частота встречаемости юридических терминов в двадцати (20) русских текстах ООН, могут быть использованы в качестве предикторов для определения сложности русских текстов. Результаты показали, что большинство текстов являются плотными, то есть все они превышают 40%, что, согласно теории Ure, если текст превышает 40%, он, скорее всего, предоставляет много информации, и чем больше информации в тексте, тем больше понимания он нам дает. На основе двадцати письменных текстов, которые мы проанализировали, исследование доказывает, что как лексическая плотность, так и разнообразие коррелируют со сложностью текста, поскольку соотношение типов лексем (TTR) в выбранных текстах варьируется от 0,56 до 0,62, что подразумевает, что тексты являются как лексически плотными, так и разнообразными. Уровень плоти Кинкейда, основанный на длине предложения, количестве слов, количестве слогов, оказался относительно высоким. Например, RT16 (11,75) – самый легкий для обработки текст из всех текстов, однако RT13 (16,49) – самый трудный для чтения текст, что отвечает на вышеупомянутый вопрос исследования – *каковы лексические типы русских текстов ООН?*

После определения лексической плотности в данном исследовании мы обнаружили, что большинство текстов плотные, что означает, что тексты информативны из-за наличия большого количества лексических единиц, которые передают

смысл читателю. Лексическое разнообразие, которое соотносится с TTR, показало, что тексты варьируются от 0,56 до 0,62, что означает, что русские тексты ООН схожи в плане лексического разнообразия. Можно также сказать, что уровень плоти Кинкейда показывает, что большинство русских текстов ООН попали в 16 баллов, что является самым высоким показателем в данном исследовании, поэтому эти тексты трудны для обработки. Основываясь на результатах, полученных с помощью уровень плоти Кинкейда, мы можем прийти к выводу, что этот показатель является индикатором сложности (трудности) текста. Как уже отмечалось в данном исследовании, большое внимание уделяется также частотности юридических терминов, поскольку корпус основан на русских текстах ООН. Многие тексты содержат большее количество падежей юридических терминов; возможно, автор сделал это для облегчения понимания. Таким образом, мы можем предположить, что частота встречаемости юридических терминов в юридическом тексте, скорее всего, делает текст либо сложнее, либо проще. То есть, если в тексте повторяются одни и те же юридические термины, читатель, скорее всего, поймет весь смысл данного текста. Основные достижения, включая вклад в данную область (оценка сложности текста), можно суммировать следующим образом: полученные результаты могут быть использованы в образовании, изучении сложности текста и переводе. Для будущих исследований мы хотели бы сравнить и сопоставить не только сложность русских и английских текстов ООН, но и сложность российских учебников по обществознанию. Такие аспекты текста, как частота слов, синтаксическая сложность и многие другие показатели были непомерно сложны для вычисления 30 лет назад, но теперь их можно вычислить без особых усилий.

Литература

1. Biber D. Methodological issues regarding corpus-based analyses of linguistic variation // Literary and Linguistic Computing. 1990. Vol. 5. Iiss. 4. P. 257 – 269.

2. Biber D. On the complexity of discourse complexity: a multidimensional analysis // Discourse Processes. 1992. Vol. 15. Iss. 2. P. 133 – 163.

3. Biber D. Variation across speech and writing. Cambridge: Cambridge Univ. Press, 1988. 299 p.

4. Castello E. Tourist-information texts: a corpus-based study of four related genres. Wallingford: CABI, 2002. P. 207 – 216.

5. Daller M.H., Hout van R., Treffers-Daller J. Lexical richness in the spontaneous speech of bilinguals // Applied Linguistics. 2003. № 24. P. 197 – 222.

6. DuBay W.H. The principles of readability. Costa Mesa: Impact Information, 2004. 74 p.

7. Eggins S. An introduction to systemic functional linguistics. 2nd ed. New York: Continuum, 2004. 384 p.

8. Fang Z. The language demands of science reading in Middle School // International Journal of Science Education. 2006. № 28. P. 491 – 520.

9. Gayraud F. Le développement de la différenciation oral/ecrit vu à travers le lexique: Ph. D. diss. Lyon: Université Lumière, 2000.

10. Gerot L., Wignell P. Making sense of functional grammar. New south wales: antipodean educational enterprises, 1994. 258 p.

11. Goulden R., Nation P., Read J. How large can a receptive vocabulary be? // Applied Linguistics. 1990. Vol 11. № 4. P. 341 – 363.

12. Halliday M.A.K. Spoken and written language. Geelong Victoria: Deakin Univ. Press, 1985. 109 p.

13. Halliday M.A.K. Complementarities in language. Beijing: The Commercial Press, 2008. 229 p.

14. Halliday M A.K., Webster J.J. Methods-techniques-problems // Continuum companion to systemic functional linguistics. London: Continuum, 2009. P. 59 – 86.

15. Halliday M.A.K., Martin J.R. Some Grammatical Problems in Scientific English // Writing science: literacy and discursive power. London: The Falmer Press, 1993. P. 76 – 94.

16. Halliday M.A.K. Spoken and written modes of meaning // Media texts, authors and readers: a reader / ed. D. Graddol, O. Boyd-Barrett. Clevedon [England]; Philadelphia: Multilingual Matters in association with The Open University, 1994. P. 51 – 73.

17. Halliday M.A.K., rev. Matthiessen C.M. I.M. An introduction to functional grammar. 3rd ed. London: Hodder Arnold, 2004. 688 p.

18. Halliday M.A.K., Hasan R., Fawcett R.P. The grammarian's dream: lexis as most delicate grammar // New developments in systemic linguistics: theory and description. London: New York: Frances Pinter, 1987. P. 184 – 211.

19. Hirsh D., Nation P. What vocabulary size is needed to read unsimplified texts for pleasure? // Reading in a foreign language. 1992. Vol. 8. P. 689 – 696.

20. Pentimonti J.M., Zucker T.A., Justice L.M. et al. Informational text use in preschool classroom read-alouds // The reading teacher. 2010. Vol. 63. P. 656 – 665.

21. Johansson V. Lexical diversity and lexical density in speech and writing: a developmental prespective // Working papers. 2008. Vol. 53. P. 61 – 79.

22. Klare G.R. Assessing readability // Reading research. 1974. Quart. 10. P. 62 – 102.

23. Lassen I. Accessibility and acceptability in Technical Manuals: a survey of style and grammatical metaphor. Amsterdam: John Benjamins, 2003. 183 p.

24. Malvern D., Richards B., Chipere N. et al. Lexical diversity and language development: quantification and assessment. Basingstoke, UK: Palgrave Macmillan, 2004. 288 p.

25. Merlini-Barbaresi L. Complexity in language and text. Pisa: Edizioni Plus (Università di Pisa), 2003. 468 p.

26. Nation P., Waring R. Vocabulary size, text coverage and word lists // Vocabulary: description, acquisition, and pedagogy / ed. N. Schmitt, M. McCarthy. Cambridge: Cambridge Univ. Press, 1997. P. 6 – 19.

27. O'Donnel M. Introduction to systemic functional linguistics for discourse analysis. Madrid: Autonomous University of Madrid, 2012.

28. Sauro S., Smith B. Investigating L2 performance in text chat // Applied linguistics. 2010. Vol. 31. № 4. P. 554 – 577.

29. Solovyev V., Ivanov V., Solnyshkina M. Assessment of reading difficulty levels in russian academic texts: approaches and metrics // Joutnal of intelligent and Fuzzy Sistems. 2018. № 34. P. 1 – 10.

30. Stern G. Learners' grammar dictionary. Singapore: Learner Publishing Pte Ltd, 2000. 234 p.

31. Ure J. Lexical density and register differentiation // Applications of linguistics: selected papers of the Second International Congress of Applied Linguistics (Cambridge 1969) / ed. G.E. Perren, J.L. M. Trim. Cambridge: Cambridge Univ. Press, 1971. P. 443 – 452.

32. Ure J., Ellis J. Register in descriptive linguistics and linguistic sociology // Issues in sociolinguistics / ed O. Uribe-Villegas. The Hague: Mouton, 1977. P. 197 – 244.

33. Wolfe-Quintero K., Shunju I., Hae-Young K. Quantifying lexical diversity in the study of language development. Hawaii: Second Language Teaching and Curriculum Center. University of Hawaii, 1998.

34. Yu H., Wu C. Text complexity as an indicator of translational style: a case study // Linguistics and the Human Sciences. 2018. Vol. 13. P. 179 – 200.

35. Zipf G.K. Human behavior and the principle of least effort. Reading, MA: Addison-Wesley, 1949. 600 p.

References

1. Biber D. Methodological issues regarding corpus-based analyses of linguistic variation. Literary and Linguistic Computing. 1990. Vol. 5. Iiss. 4. P. 257 – 269.

2. Biber D. On the complexity of discourse complexity: a multidimensional analysis. Discourse Processes. 1992. Vol. 15. Iss. 2. P. 133 – 163.

3. Biber D. Variation across speech and writing. Cambridge: Cambridge Univ. Press, 1988. 299 p.

4. Castello E. Tourist-information texts: a corpus-based study of four related genres. Wallingford: CABI, 2002. R. 207 – 216.

5. Daller M.H., Hout van R., Treffers-Daller J. Lexical richness in the spontaneous speech of bilinguals. Applied Linguistics. 2003. № 24. P. 197 – 222.

6. DuBay W.H. The principles of readability. Costa Mesa: Impact Information, 2004. 74 p.

7. Eggins S. An introduction to systemic functional linguistics. 2nd ed. New York: Continuum, 2004. 384 p.

8. Fang Z. The language demands of science reading in Middle School. International Journal of Science Education. 2006. № 28. P. 491 – 520.

9. Gayraud F. Le développement de la différenciation oral/ecrit vu à travers le lexique: Ph. D. diss. Lyon: Université Lumière, 2000.

10. Gerot L., Wignell P. Making sense of functional grammar. New south wales: antipodean educational enterprises, 1994. 258 p.

11. Goulden R., Nation P., Read J. How large can a receptive vocabulary be? Applied Linguistics. 1990. Vol 11. № 4. P. 341 – 363.

12. Halliday M.A.K. Spoken and written language. Geelong Victoria: Deakin Univ. Press, 1985. 109 p.

13. Halliday M.A.K. Complementarities in language. Beijing: The Commercial Press, 2008. 229 p.

14. Halliday M A.K., Webster J.J. Methods-techniques-problems. Continuum companion to systemic functional linguistics. London: Continuum, 2009. P. 59 – 86.

15. Halliday M.A.K., Martin J.R. Some Grammatical Problems in Scientific English. Writing science: literacy and discursive power. London: The Falmer Press, 1993. P. 76 – 94.

16. Halliday M.A.K. Spoken and written modes of meaning. Media texts, authors and readers: a reader.ed. D. Graddol, O. Boyd-Barrett. Clevedon [England]; Philadelphia: Multilingual Matters in association with The Open University, 1994. P. 51 – 73.

17. Halliday M.A.K., rev. Matthiessen C.M. I.M. An introduction to functional grammar. 3rd ed. London: Hodder Arnold, 2004. 688 p.

18. Halliday M.A.K., Hasan R., Fawcett R.P. The grammarian's dream: lexis as most delicate grammar. New developments in systemic linguistics: theory and description. London: New York: Frances Pinter, 1987. P. 184 – 211.

19. Hirsh D., Nation P. What vocabulary size is needed to read unsimplified texts for pleasure? Reading in a foreign language. 1992. Vol. 8. P. 689 – 696.

20. Pentimonti J.M., Zucker T.A., Justice L.M. et al. Informational text use in preschool classroom readalouds. The reading teacher. 2010. Vol. 63. P. 656 – 665.

21. Johansson V. Lexical diversity and lexical density in speech and writing: a developmental prespective. Working papers. 2008. Vol. 53. P. 61 – 79.

22. Klare G.R. Assessing readability. Reading research. 1974. Quart. 10. P. 62 – 102.

23. Lassen I. Accessibility and acceptability in Technical Manuals: a survey of style and grammatical metaphor. Amsterdam: John Benjamins, 2003. 183 p.

24. Malvern D., Richards B., Chipere N. et al. Lexical diversity and language development: quantification and assessment. Basingstoke, UK: Palgrave Macmillan, 2004. 288 p.

25. Merlini-Barbaresi L. Complexity in language and text. Pisa: Edizioni Plus (Università di Pisa), 2003. 468 p.

26. Nation P., Waring R. Vocabulary size, text coverage and word lists. Vocabulary: description, acqui-sition, and pedagogy. ed. N. Schmitt, M. McCarthy. Cambridge: Cambridge Univ. Press, 1997. P. 6 – 19.

27. O'Donnel M. Introduction to systemic functional linguistics for discourse analysis. Madrid: Autonomous University of Madrid, 2012.

28. Sauro S., Smith B. Investigating L2 performance in text chat. Applied linguistics. 2010. Vol. 31. № 4. P. 554 – 577.

29. Solovyev V., Ivanov V., Solnyshkina M. Assessment of reading difficulty levels in russian academic texts: approaches and metrics. Joutnal of intelligent and Fuzzy Sistems. 2018. № 34. P. 1 – 10.

30. Stern G. Learners' grammar dictionary. Singapore: Learner Publishing Pte Ltd, 2000. 234 p.

31. Ure J. Lexical density and register differentiation. Applications of linguistics: selected papers of the Second International Congress of Applied Linguistics (Cambridge 1969). ed. G.E. Perren, J.L. M. Trim. Cambridge: Cambridge Univ. Press, 1971. P. 443 – 452.

32. Ure J., Ellis J. Register in descriptive linguistics and linguistic sociology. Issues in sociolinguistics. ed O. Uribe-Villegas. The Hague: Mouton, 1977. P. 197 – 244.

33. Wolfe-Quintero K., Shunju I., Hae-Young K. Quantifying lexical diversity in the study of language development. Hawaii: Second Language Teaching and Curriculum Center. University of Hawaii, 1998.

34. Yu H., Wu C. Text complexity as an indicator of translational style: a case study. Linguistics and the Human Sciences. 2018. Vol. 13. P. 179 – 200.

35. Zipf G.K. Human behavior and the principle of least effort. Reading, MA: Addison-Wesley, 1949. 600 p.

***

Comparative study of Russian UN texts

Mariko Mohamed Lamin, Postgraduate,
Kazan (Volga Region) Federal University

Abstract: our research is aimed at building a lexical typology of Russian UN texts by defining the range of metrics of a set of parameters i.e., Flesh-Kincaid Grade Level (FKGL), lexical density, lexical diversity (TTR) and Legal terms incidence.We utilized the data from the UN site ((https://www.ohchr.org/EN/PublicationsResources/Pages/RecentPublications.aspx) and compiled a corpus of 20 UN texts with the total size of about 20,000 words. FKGL, which is computed based on an average sentence length, number of words, number of syllables, proves to be relatively high and varies between 11.75 and 16.49. The findings indicate that the texts are both lexically dense and diverse: with lexical density varying from 67 to 74 and TTR ranging from 0.56 to 0.62. Thus, they demonstrate a high degree of correlation with FKGL. A typical Russian UN text contains about 66 legal terms, which shows that 7 texts out 8 contain legal terms incidence varying from 66 to 73. We conclude that the range of metrics of the four parameters selected above could serve as predictors in determing complexity in Russian UN texts thus confirming Ure's views (1971). The results can be of interest to educators, and researchers of text complexity and translation studies.

Keywords: Flesh Kincaid Grade Level or text complexity, lexical density, lexical diversity (TTR), legal terms incidence, Russian UN texts

For citation: Mariko Mohamed Lamin Comparative study of Russian UN texts. Modern Humanities Success. 2023. 8. P. 57 – 69.

Received: May 28, 2023; Revised: June 20, 2023; Accepted: July 31, 2023.