DOI: https://doi.org/10.31992/0869-3617-2020-29-7-89-103

Изучение англоязычного академического письма инструментами компьютерной лингвистики

Шпит Елена Ирисметовна – ст. преподаватель кафедры иностранных языков. E-mail: forester_2007@mail.ru
Томский государственный университет систем управления и радиоэлектроники, Томск, Россия
Адрес: 634050, Томская область, г. Томск, проспект Ленина, 40
Куровский Василий Николаевич – д-р пед. наук, проф. E-mail: v.kurovskii@yandex.ru
Томский государственный педагогический университет, Томск, Россия
Адрес: 634061, Томская область, г. Томск, ул. Киевская, 60

Аннотация. *Написание научного текста на английском языке молодым учёным, только начинающим свою публикационную деятельность, сопровождается определёнными трудностями, связанными с переводом стилистически ярко окрашенных предложений с русского языка. Изучение особенностей любого стиля невозможно без анализа образцов дискурса, что актуализирует использование компьютерной лингвистики, поскольку она позволяет автоматизировать многие механизмы обработки языковых и текстовых материалов и производит достаточно точные количественные данные. Данное исследование рассматривает применение программ AntConc и Coh-Metrix для проведения сравнительного анализа аннотаций магистрантов к научным статьям, написанным для дальнейшей публикации в международных журналах (ученический корпус), и аннотаций к научным статьям исследователей из разных стран мира, уже опубликованным в высокорейтинговых англоязычных журналах (контрольный корпус). Анализ корпусов в упомянутых ресурсах позволил выявить несовершенства и достоинства студенческих аннотаций, охарактеризовать их на уровне лексики, синтаксиса и дискурса, а также обозначить перспективы использования указанных программ в обучении навыкам академического письма.*

Ключевые слова: *академическое письмо, научный текст, компьютерная лингвистика, автоматизированный анализ текста, AntConc, Coh-Metrix, специализированный корпус*

Для цитирования: *Шпит Е.И., Куровский В.Н.* Изучение англоязычного академического письма инструментами компьютерной лингвистики // Высшее образование в России. 2020. Т. 29. № 7. С. 89-103.
DOI: https://doi.org/10.31992/0869-3617-2020-29-7-89-103

Введение

Научный текст, создаваемый носителями разных языков, в основе своей имеет схожие признаки. В первую очередь, это цель повествования – познакомить научное сообщество с результатами своей исследовательской работы. Далее – средства, которые применяются для связного, логичного и убедительного выстраивания своих мыслей, чтобы информация была воспринята корректно и недвусмысленно. Для текстов естественно-научных и научно-техниче-

ских подстилей, представляющих для нас особый интерес, характерны «логичность, отвлечённость, стандартизированность формы, информационная ёмкость, нейтральная модальность» [1]. Общими внешними языковыми признаками можно считать следующие: предпочтение глагольных форм настоящего времени и пассивного залога, высокая частотность абстрактных отглагольных и отадъективных существительных, широкое употребление терминологии и групп существительных в родительном падеже, интернациональных слов и слов латинского и греческого происхождения и т.д. Однако в силу различных морфологических и синтаксических особенностей русского и английского языков возникает ряд проблем, вызывающих у автора сложности с переводом на английский язык, а у читателя – с восприятием переведённого текста.

Научный стиль русского языка накладывает яркий отпечаток на язык научных текстов в переводном варианте и является предметом изучения многими лингвистами, методистами и педагогами. «Межъязыковая интерференция рассматривается исследователями как наиболее мощный фактор отрицательного воздействия родного языка на изучаемый иностранный, так как практика преподавания свидетельствует о том, что наиболее стойкими оказываются ошибки, вызванные интерферирующим влиянием системы родного языка, укоренившегося в сознании обучающихся» [2]. В работе автор приводит множество примеров межъязыковой интерференции в работах студентов разных стран мира. Вопросам коммуникативно-значимых (нарушение порядка слов в английском предложении, ошибки в использовании активного и пассивного залогов, отсутствие подлежащего или сказуемого в предложении др.) и коммуникативно-незначимых (связанных с различиями в синтаксической организации письменной речи и др.) ошибок русскоязычных авторов посвящена работа О.Л. Добрыниной [3]. В [4] она рассматривает снижение «удобочитаемости» текстов, вызванное стилистическими погрешностями, такими как, например, излишняя номинализация и злоупотребление пассивным залогом, и даёт советы по предварительной подготовке текста к переводу на английский язык. Эффекты использования пассивного залога и деперсонализации «на русский манер» отражены в работе [5]. Наше исследование [6] выявило сильное влияние интерференции на специфические языковые аспекты (атрибутивные группы, пунктуация, артикли, комментирование таблиц и графиков и др.) у магистрантов Томского государственного университета систем управления и радиоэлектроники (ТУСУР).

В контексте глобализации и интернационализации научного общения вопрос читабельности текста приобретает всё большее значение. А. Волворк, автор 20 книг по академическому и профессиональному английскому языку, указывает на различия в отношении к читателю в разных культурах [7]. Например, в восточной риторике ответственность за корректное восприятие связей между предложениями, абзацами и информацией в тексте лежит на читателе. В традициях английской риторики, которая стала нормой международного академического взаимодействия, эта ответственность лежит на авторе, а читателю остаётся быстро и адекватно воспринимать подаваемую информацию, прикладывая минимум усилий. Такой подход, по мнению автора, обеспечивает гораздо более широкую читательскую аудиторию и, как результат, высокую вероятность цитируемости. О приоритете понимаемости сути текста ещё в прошлом веке говорил известный филолог, культуролог и искусствовед Д.С. Лихачёв: «Внимание читающего должно быть сосредоточено на мысли автора, а не на разгадке того, что автор хотел сказать» [8].

Наша статья посвящена изучению 37 аннотаций к научным статьям, написанных молодыми учёными, для публикации в меж-

дународных журналах и трудах конференций (ученический корпус). Для сравнения был составлен специализированный корпус из 42 аннотаций к статьям, уже опубликованным в рецензируемых международных журналах и материалах конференций (контрольный корпус). Изучение аннотаций не случайно; это второй по обращению, после заголовка, элемент статьи и, следовательно, должен получать не меньшее, а, возможно, большее внимание со стороны автора. Аннотации должны выполнять важные функции: как можно более компактно, но полно передавать основное содержание статьи; делать это наиболее понятным и корректным языком и «продавать статью» потенциальным читателям путём привлечения их внимания, т.е. побуждать их прочесть статью целиком [7]. Кроме того, данный компонент статьи очень подходит для текстового анализа из-за сконцентрированности языковых и стилистических явлений в относительно законченном тексте.

Целью данного исследования является верификация гипотезы о том, что глубокая языковая интерференция ведёт к снижению когезии и когерентности научного текста. Кроме того, мы определим основные характеристики англоязычных аннотаций молодых русскоязычных авторов в сравнении с аннотациями англоговорящих авторов. Исследование будет построено на анализе трёх уровней текстового проявления: слово, предложение и дискурс.

Методы исследования. Используемые корпусы

Ученический корпус. Представляет собой набор из 37 аннотаций (4085 словоупотреблений – токенов) из научных статей, написанных молодыми учёными кафедры телевидения и управления ТУСУР в области электромагнитной совместимости для публикации в международных журналах и материалах международных конференций. Статьи представляют собой первичные научные тексты, написанные в соответствии с требованиями *IEEE*[1], которые предъявляются к рукописям, подаваемым в издания этой организации разного уровня. Чтобы соблюсти полную объективность исследования, используемые статьи не были редактированы, т.е. никакие ошибки, включая грамматические и орфографические, не были исправлены. Известно, что внесение некоторых корректив со стороны редактора может повлечь за собой изменение целого предложения или параграфа, что, в свою очередь, ведёт к нарушению оригинальности текста и коренным изменениям в результатах цифровой обработки. Авторы проводят свои исследования под научным руководством ведущих учёных вуза, большинство имеют некоторый опыт написания статей на английском языке и достаточно уверенно владеют специальной терминологией. Предположительно, они умеют пользоваться интернет-ресурсами для данных целей.

Контрольный корпус. Составлен из 42 аннотаций (5480 токенов), извлечённых из научных статей, написанных авторитетными учёными и опубликованных в высокорейтинговых международных изданиях. При отборе статей для данного корпуса необходимо было соблюсти условие репрезентативности. В данном случае репрезентативность выражается в следующих факторах: 1) первоначально статьи были отобраны профилирующей кафедрой из источников, которыми пользуются исследователи кафедры; 2) тексты, написанные русскоязычными авторами, не входили в число отобранных статей, чтобы максимально исключить влияние интерференции русского языка и стиля; 3) включались статьи, написанные авторами из США и Великобритании, а также из других стран мира (Китая, Германии, Швеции, Канады и др.), поскольку вопросами электромагнитной

[1] IEEE Author Centre Journals, Templates for transactions. URL: https://journals.ieeeauthorcenter.ieee.org/create-your-ieee-journal-article/authoring-tools-and-templates/ieee-article-templates/templates-for-transactions/

совместимости (ЭМС) занимаются академические сообщества в различных странах; тем самым они вносят определённый вклад в языковые и стилистические особенности данного дискурса.

Обработка корпусов

Оба корпуса были изучены с помощью двух открытых онлайн-программ: *AncConc* (https://www.laurenceanthony.net/software/antconc/) и *Coh-Metrix (3.0)* (www.cohmetrix.com), а именно, *Coh-Metrix Web Tool*.

AntConc – это бесплатная многоплатформенная программа-конкордансер, созданная Э. Лоуренсом, профессором Университета г. Васеда (Япония), в качестве инструмента для автоматизированного анализа текстов. Программа предлагает удобный графический пользовательский интерфейс с мощным инструментарием отображения сочетаемости слов, генератором частотности, функциями анализа кластеров и лексических словосочетаний, а также с графиком дистрибуции слова в тексте [9]. Возможности данного конкордансера в сравнении с другими подобными ресурсами, а также актуальность создания и применения специализированных корпусов в современном информационном сообществе описаны в [10]. В [11] создатель *AntConc* совместно с коллегами из Университета г. Васеда демонстрирует опыт использования программы при анализе исследовательских статей по математике на макро- (следование структуре IMRaD) и микро- (стиль письма в текстах данного направления научного знания) уровнях. В [12] приводятся примеры использования *AntConc* в индуктивном обучении (т.е. обучении, при котором знания создаются учащимися в процессе изучения образцов действительного мира, в данном случае – образцов речи, представленных различными по типу, жанру и литературности текстами). В.Э. Рогачева в [13] использует *AntConc* для анализа и сопоставления данных, извлечённых из двуязычного текстового корпуса с целью установления переводных эквивалентов между терминами в единой предметной области. В представляемом исследовании ресурс был использован для того, чтобы выявить сходства и различия языкового материала аннотаций к статьям по ЭМС с помощью анализа частотности использования некоторых лексических и грамматических структур.

Coh-Metrix был создан А. Грейсером и Д. МакНамара в Университете г. Мэмфис (США). Он представляет собой один из инструментов цифровой лингвистики, который продуцирует индексы лингвистического и дискурсивного представления текста. Создаваемые индексы могут быть использованы для проведения разного рода исследований когезии (cohesion) текста и когерентности (coherence) его восприятия. Текущая онлайн-версия *Coh-Metrix 3.0* измеряет тексты объёмом до 15000 символов по 108 критериям. Самыми яркими примерами использования *Coh-Metrix* являются исследования Ф. МакКарти и его соавторов [14], в которых были выявлены отличия письма *японских* учёных от письма их *британских* и *американских* коллег. Вслед за ними и используя их находки, Б. Данкин и Ч. Холл [15] обнаружили значительные лингвистические и дискурсивные различия между аннотациями *американских* и *корейских* учёных по биомедицине, проявившиеся в значительно меньшем лексическом и синтаксическом разнообразии последних и в других аспектах. *Coh-Metrix* был использован в [16] для измерения сложности научных текстов, созданных в образовательных целях, на основе двух независимых факторов: референциальное пересечение (referential overlap) и доступность лексического наполнения текстов (vocabulary accessibility). Разнообразие областей применения *Coh-Metrix* отражено в работе [17], оно варьируется от дифференцирования и отбора учебной литературы до формирования различных вариантов образовательной среды в целом. В российской прикладной лингвистике *Coh-Metrix* позволил обнаружить, что тексты ЕГЭ по английскому языку «характеризуются более

высокими показателями конкретности и более низкими показателями синтаксической простоты, чем тексты FCE» [18].

В нашем исследовании ученический корпус сопоставляется с контрольным корпусом *по трём уровням текстового анализа*: слово (лексическое разнообразие, частотность использования отдельных частей речи, полисемия и гиперонимия), предложение (количество слов перед основным глаголом; использование активного/пассивного залогов и служебных слов) и дискурс (сложность текста с точки зрения синтаксиса, семантики слов и когезии).

Результаты и обсуждение

Сравнительный анализ двух корпусов включает данные, полученные различными способами. В первую очередь, это данные, полученные путём измерений, проведённых в программах *AntConc* и *Coh-Metrix*. Кроме того, иногда наши результаты будут сопоставлены с результатами, полученными другими исследователями – П. МакКарти и коллегами [14] (ЯУ – японские учёные, АБ – американские и британские соответственно) и Б. Данкином и Ч. Холлом [15] (КК – корейские учёные для корейских журналов, АК – корейские учёные для американских журналов, АА – американские учёные для американских журналов). Мы также приведём нормы *Coh-Metrix* для университетского уровня (Norm College Level – Science [19]), поскольку считаем, что это уровень, который может стать точкой опоры по многим критериям для студентов, стремящихся усовершенствовать свои навыки в академическом письме.

Сравнительный анализ на уровне «СЛОВО»

Индекс «Лексическое разнообразие» измеряет диапазон вокабуляра, используемого автором: чем больше значение, тем богаче словарный запас. Этот критерий измеряется различными способами, самый простой из которых – Type-Token Ratio (TTR): отношение уникальных, неповторяющихся, слов к общему количеству токенов. Однако при таком подходе с увеличением объёма текста уменьшается соотношение. В нашем исследовании мы также приводим результаты другого подхода – VOCD (vocabulary density), который совмещает данные TTR и многократной произвольной выборки для создания коэффициента D (10–100) [20]. Из *таблицы 1* видно, что при относительно одинаковом значении TTR результаты по VOCD кардинально различны. Более того, этот показатель значительно уступает университетской норме. Исследуемые авторы из [14] также выигрывают: 76,45 (ЯУ), 83,09 и 90,26 (АБ).

В качестве примера приведём данные о разнообразии глаголов в наших корпусах, полученные с помощью *AntConc*. Количество типов глаголов в контрольном корпусе составляет 22,4 (на 1000 токенов), а в ученическом корпусе – 14,2, что почти в 1,5 раза меньше. В *Coh-Metrix* этот показатель выражен в количестве глаголов на 1000 словоупотреблений и также имеет более низкое значение в ученическом корпусе. Подобная ситуация наблюдается и с прилагательными.

Количество существительных, напротив, демонстрирует приблизительное равенство их употребления. Если сравнить полученные индексы с нормой для работ университетского уровня, то можно заметить ярко выраженные лексические особенности русского академического дискурса, а именно большую насыщенность текста существительными и ограниченное употребление глаголов (почти в три раза), т.е. номинативность речи. Примечательно отношение авторов учебного пособия к номинализации: «Именной характер научной речи не означает её содержательного упрощения. Отглагольное существительное наряду с другими средствами языка способно нести дополнительные, свёрнутые смыслы, то есть выражать полипропозитивность. Под полипропозитивностью понимается семантико-синтаксическая *усложнённость простого предложения, его способность выражать*

Таблица 1

Лексический анализ
(ЛД – лексическое разнообразие, N – количество, СЗ – среднее значение, Т – токен)

Table 1

Word-level analysis

№	Параметр	Ученич. корпус	Контр. корпус	Универ. норма
1	**Lexical diversity**, TTR, content words lemmas (ЛД, TTR, знаменательные слова, по шкале 0 – 1)	0,811	0,822	0,693
2	**Lexical diversity**, VOCD, all words (ЛД, VOCD, коэффициент D)	49,260	80,778	76,040
3	**Verb incidence** (N глаголов на 1000 Т)	101,959	114,093	111,054
4	**Adjective incidence** (N прилагательных на 1000 Т)	107,635	127,279	98,167
5	**Noun incidence** (N существительных на 1000 Т)	326,604	321,851	290,676
6	**Gerund density**, incidence (N герундиев на 1000 Т)	22,683	21,886	6,366
7	**Infinitive density**, incidence (N инфинитивов на 1000 Т)	3,896	15,124	6,026
8	**Polysemy** for content words, mean (многозначность знаменательных слов, СЗ)	3,734	3,646	3,929
9	**Hypernymy** for nouns, mean (гиперонимия существительных, СЗ)	6,814	6,519	6,397

не один, а несколько элементарных смыслов, соответствующих актуальным фактам действительности» [1].

Использование герундиев (формы глагола, которая часто выступает в роли существительного в русском языке) считается признаком высокого уровня владения языком и зачастую избегается не-англоговорящими авторами либо по незнанию, либо во избежание некорректного употребления [14; 15]. Однако в наших результатах их количество практически одинаково и значительно выше университетской нормы. Более того, данные по нашим корпусам выше данных в [14]: 10,58 (ЯУ), 16,11 и 17,22 (АБ). Ситуация подсказывает, что в научном сообществе по ЭМС использование герундиев является нормой, а наши студенты, опираясь в качестве источников на статьи зарубежных авторов, осознанно или неосознанно перенимают эти нормы. Это в очередной раз доказывает необходимость исследований в области дисциплинарных и узкодисциплинарных научных текстов как для лингвистических, так и для педагогических целей.

Инфинитивы представляют собой отдельную довольно сложную тему в силу невероятно высокой номинализации научного стиля в русском языке. Там, где англоязычные авторы использовали бы инфинитив цели (to simulate), русскоязычные авторы обязательно используют существительное (for the simulation of). Данное отличие двух способов выражения цели приводит к увеличению количества слов в предложении, а также к ещё большей номинализации текста; и то и другое неизбежно затрудняет понимание текста [1]. В нашем случае количество используемых студентами инфинитивов почти в четыре раза меньше по сравнению с контрольным корпусом. В сравнении с [14; 15] ситуация аналогичная: 14,22 (ЯУ), 18,81 и 19,93 (АБ) в [14]; 9,681 (КК), 11,261 и 13,238 (АК и АА соответственно) в [15].

Таблица 2

Синтаксический анализ
(N – количество, СЗ – среднее значение, Т – токен)

Table 2

Sentence-level analysis

№	Параметр	Ученич. корпус	Контр. корпус	Универ. норма
1	**Sentence length**, number of words, mean (длина предложений, СЗ)	21,7764	23,781	17,715
2	**Left embeddedness**, words before main verb, mean (N слов перед основным глаголом, СЗ)	8,533	6,328	5,070
3	***Of*-phrase density**, incidence (*AntConc*) (N *of*-фраз на 1000 Т)	79,31	42,01	-
4	**Agentless passive voice density**, incidence (N глаголов в пассивном залоге на 1000 Т)	41,221	24,432	8,914
5	**Passive voice at the end**, incidence (*AntConc*) (N пассивных форм в конце предложения, 1000 Т)	18,4	4,2	-
6	**Active Voice**, incidence (*AntConc*) (N глаголов в активном залоге, на 1000 Т)	17,6	36,1	-
7	**First person plural pronoun**, incidence (N «We» на 1000 Т)	0,655	2,772	4,361
8	**All connectives**, incidence (N служебных слов на 1000 Т)	59,547	80,20	82,993

Чтобы понять семантику использованных слов, мы выбрали такие индексы, как полисемия и гиперонимия слова, посредством которых *Coh-Metrix* измеряет многозначность и абстрактность/конкретность слов. Эти измерения базируются на онлайн-инструменте *WordNet* [21], который группирует слова в наборы синонимов, связанных семантическими отношениями. Высокое значение полисемии означает многозначность слова. Гиперонимия относится к количеству смысловых уровней, которое слово имеет выше в концептуальной, таксономической иерархии. Слова с низким значением гиперонимии – абстрактные слова, поскольку не имеют, как правило, конкретных значений. Результаты измерений дают понять, что данные показатели в обоих корпусах вполне сопоставимы и свидетельствуют о том, что используемая лексика отличается высокой степенью конкретности и однозначности. Такая лексика не представляет больших трудностей для осведомлённого читателя. Полисемия измерялась в [14] и составила 2,78 (ЯУ) и 2,83 и 3,01 (АБ). В [15] была измерена гиперонимия, она составила 4,421 (КК) и 4,080 и 4,404 (АК и АА соответственно). Относительно равные параллельные значения этих критериев дают понимание присущей данной отрасли науки лексики.

Сравнительный анализ на уровне «ПРЕДЛОЖЕНИЕ»

Синтаксический анализ обнаружил гораздо больше различий, чем лексический. При относительно одинаковой длине предложений строение предложений в ученическом корпусе отличается более сложным синтаксисом (*Табл. 2*).

Количество слов перед основным глаголом значительно больше в ученических аннотациях, причём 21 из 37 аннотаций имеют значение выше среднего. Избыточное число слов перед сказуемым требует усиленной работы памяти и тем самым затрудняет про-

цесс восприятия информации. Для сравнения: в [15] этот индекс равен 6,812 для АА; это значит, что необходимо перестраивать предложение, а возможно, всю мысль в предложении, чтобы сделать его более читабельным и более сопоставимым с принятыми нормами в международном научном сообществе.

Большое количество слов перед сказуемым тесно связано с использованием *of*-фраз. Предлог «of» используется в образовании родительного падежа, который является способом обозначения определительных отношений (the conductors of a PCB), а также объектных отношений, возникающих в сочетании с отглагольными именами существительными (the development of a modal filter) [1]. Результаты измерений корпусов в *AntConc* показали, что частота использования *of*-фраз в ученическом корпусе почти в два раза выше, чем в контрольном корпусе. Это можно объяснить несколькими причинами: неумением или страхом создавать крупные атрибутивные фразы, заменой инфинитивов цели и герундиев существительными с предлогом (for the improvement of; by the use of) или неправильным употреблением герундия (by using of a meander line). Такие огрехи типичны для начинающих авторов и потому должны быть рассмотрены в процессе обучения академическому письму на английском языке.

Студенческие тексты также отличаются гораздо большим (почти в два раза) количеством глаголов в пассивном залоге, что вызвано деагентивностью русского научного стиля. Это может привести к снижению когерентности, особенно если сказуемое в пассиве завершает цепочку слов в предложении. Результаты, полученные с помощью *AntConc*, показывают, что количество глаголов в пассиве в конце предложений в ученическом корпусе в 4,4 раза выше, чем в контрольном. Использование исключительно страдательного залога для перевода обобщённо-личных и безличных конструкций приводит к нескольким отрицательным результатам: 1) аннотация выглядит как набор предложений, а не связный текст («microcosm» [7]); 2) количество слов, расположенных до сказуемого, зачастую значительно превышает значения 8–10, рекомендуемые англоязычными специалистами в области академического письма [7]; 3) мысль предложения теряется по мере прохождения через все составляющие группы подлежащего к основному глаголу. Если автор стремится попасть на страницы высокорейтингового журнала, ему придётся пересмотреть синтаксис не только предложений по отдельности, но и продумать иное строение всего абстракта. Отметим соотношение активного и пассивного залогов в корпусах: количество предложений в активном залоге в контрольном корпусе больше в 1,5 раза, тогда как в ученическом – меньше в 2,3 раза.

С темой страдательного залога связан вопрос об использовании местоимения «мы», которое в русском языке обычно вуалируется безличными, обобщённо-личными и неопределённо-личными конструкциями [1]. Однако в английском языке использование «мы» является нормой даже в академическом дискурсе [5]. В журналах и материалах упомянутых выше конференций *IEEE*, куда часто подают свои статьи студенты исследуемого профиля, говорится, что при желании авторы могут писать от первого лица как в единственном, так и во множественном числе. Между тем «мы» в ученическом корпусе встречается всего лишь в двух аннотациях, тогда как в контрольном корпусе – в 11, что почти в четыре раза чаще.

Служебные слова (союзы, предлоги) напрямую ассоциируются с темой связности текста, поскольку создают логические связи между словами, частями и идеями предложения и формируют признаки организации текста. *Coh-Metrix* измеряет частоту использования элементов (на 1000 Т) и их характер (каузальные, логические, противопоставительные и др.). Поскольку аннотация является кратким, но законченным информативным текстом, мы рассма-

Таблица 3

Дискурсивный анализ
(СТ – сложность текста, П – перцентиль)

Table 3

Discourse-level analysis

№	Параметр	Ученич. корпус	Контр. корпус	Универ. норма
1	**Word count**, number of words (среднее кол-во слов в тексте)	110,595	128,19	287,700
2	**Text Easability PC Syntactic simplicity**, percentile (СТ: простота синтаксиса, П)	47,362 max: 96,93	32,566 max: 87,7	59,820
3	**Sentence syntax similarity**, all combinations, across paragraphs, mean (схожесть синтакс. структур в тексте, коэффициент)	0,154	0,1	0,111
4	**Text Easability PC Word concreteness**, percentile (СТ: конкретность слов, П)	39,402 max: 98,81	39,604 max: 100	50,665
5	**Text Easability PC Referential cohesion**, percentile (СТ: референциальная когезия, П)	58,2 max: 99,9	52,755 max: 98,12	61,826
6	**Text Easability PC Deep cohesion**, percentile (СТ: глубокая когезия, П)	42,469 max: 99,77	50,213 max: 100	54,898
7	**Flesch Reading Ease** (удобочитаемость по шкале Flesch)	32,378	20,297	52,164

триваем индекс частотности всех служебных слов в целом. Результаты показывают, что ученический корпус в этом плане значительно проигрывает и контрольному корпусу, и университетской норме. Можно предположить, что в большинстве аннотаций последовательность изложения является низкой.

Сравнительный анализ на уровне «ДИСКУРС»

Самой первой категорией научного текста, рассматриваемой стилистами, является его связность, или когезия. *Coh-Metrix* в своей текущей версии имеет восемь компонентов, рассматривающих связность и сложность текста в целом (Text Easability). Поскольку ресурс использует различные высокотехнологичные разработки в области вычислительной лингвистики и текстовой обработки [14], для получения общей картины на дискурсивном уровне и для верификации результатов, полученных на других уровнях анализа, мы выбрали такие аспекты, как синтаксис, лексика, референциальная и глубокая когезия. Все эти компоненты оценены по перцентилю лёгкости. Кроме того, мы включили в анализ известный критерий «оценки удобочитаемости» по Флешу[2] (*Табл. 3*).

Компонент «Простота синтаксиса» отражает степень использования простых, знакомых синтаксических структур и небольшого количества слов в предложении. Это способствует быстрому пониманию читаемого материала. По результатам измерений можно утверждать, что с этой точки зрения ученические аннотации немного более лёгкие для понимания. Если иметь в виду результаты анализа на синтаксическом уровне, можно было ожидать, что перцентиль будет более низким. Полученные данные предположительно обусловлены довольно высоким коэффициентом схожести синтаксических конструкций в данном корпусе. В [14] из-

[2] The Flesch Reading Ease and Flesch-Kincaid Grade Level. URL: https://readable.com/blog/the-flesch-reading-ease-and-flesch-kincaid-grade-level/

мерялся также индекс схожести синтаксиса; он составил 0,09 (ЯУ), 0,08 и 0,08 (АБ). В сравнении с ними видно, насколько сильно студенческие аннотации перегружены идентичными структурами.

Высокая конкретность и однозначность используемых слов и словосочетаний, выявленная в результате лексического анализа, свидетельствуют о большом количестве терминов в первичных научных статьях по результатам исследований. Они представляют определённую трудность понимания для среднестатистического человека, поэтому перцентиль лёгкости значительно ниже университетской нормы, но находится примерно на одном уровне в обоих корпусах.

Референциальная когезия оценивает пересечение слов и понятий во всём тексте. Чем больше пересечений, тем легче воспринимается информация. Результаты показывают довольно высокий и относительно схожий уровень восприятия ученических и контрольных аннотаций, близкий к университетской норме. Возможно, авторы аннотаций довольно часто прибегают к повторению одних и тех же слов, именных фраз или их сокращений.

Довольно нестабильными считаются данные по глубокой когезии для текстов небольшого объёма. Тем не менее этот индекс показывает, насколько сильны в них каузальные (since, so, because…) и интенциональные (in order to, so that, by means of…) отношения. Они помогают читающему сформировать более связное и глубокое понимание текста. Если таковых нет, то читающему приходится прилагать усилия по созданию этих связей. В ученическом корпусе по сравнению с контрольным эта когезия выражена несколько слабее и намного слабее – по сравнению с университетской нормой.

По всем трём компонентам *сложности текста*, рассматриваемым в данной работе, обращают на себя внимание максимальные значения результатов: 87,7–100. Учитывая, что анализируемые аннотации извлечены из академических текстов по довольно узкому научно-техническому профилю, такие результаты должны считаться большим минусом, означающим, что текст написан исключительно простым языком, имеет много повторений или является слишком коротким.

«Оценка удобочитаемости» по Флешу основывается на таких конкретных измерениях текста, как количество слов в предложении и количество слогов в слове. Однако при необходимости рассмотреть различные аспекты текста и дискурса или при выборе материала для учебных пособий по английскому языку опираться исключительно на этот метод измерения нельзя [22]. Результат оценки удобочитаемости по Флешу коррелирует с таблицей сложности и категорией читающих. Согласно этим таблицам аннотации из ученических корпусов считаются сложными и предназначенными для аудитории, имеющей университетское образование. Аннотации из контрольных корпусов очень сложные и также предназначены для выпускников университетов. Некоторые аннотации из обоих корпусов имеют 0 по шкале Флеша, что означает исключительно высокую степень сложности, которая может повлечь полное непонимание текста, его отклонение (рецензентами) или игнорирование целевой аудиторией.

Выводы

Результаты сравнительного анализа студенческих и контрольных аннотаций позволили количественно продемонстрировать некоторые свойства текста на английском языке, которые носят обычно интуитивный характер. Удалось сформулировать следующие особенности студенческого текста:

1) лексическое разнообразие текстов довольно низкое (более чем в два раза ниже по сравнению с контрольным корпусом и более чем в три раза – по сравнению с [14; 15]);

2) семантика используемых слов вполне сопоставима с текстами, написанными интернациональными авторами в данной области, и отличается высокой конкретностью и низкой полисемией;

3) уверенное использование герундиев вписывается в нормы исследуемого дискурсивного сообщества. Ограниченное использование инфинитива, возможно, связано с неумением сопоставлять морфологические особенности двух языков;

4) синтаксическое построение предложений заметно отличается от контрольного корпуса: преобладают пассивные формы глагола; количество слов перед основным глаголом и количество цепочек в родительном падеже (*of*-фразы) намного выше; гораздо меньше используются первое лицо во множественном числе («we») и служебные слова;

5) сложность понимания текстов, написанных русскоязычными авторами, можно объяснить отсутствием выраженной глубокой когезии и синтаксической сложностью текста, хотя высокая повторяемость синтаксических структур и референциальная когезия способствуют удобочитаемости;

6) большинство вышеобозначенных недостатков студенческих аннотаций связано с низким уровнем владения английским языком и переносом норм русского научного стиля в английский текст.

Несмотря на ограничения данного исследования, выразившиеся в относительной короткости измеряемых текстов и в невозможности учитывать грамматические и низкоуровневые ошибки (орфографию, пунктуацию), данные инструменты анализа позволяют визуализировать эффект использования тех или иных лексико-грамматических явлений, выявить основные недостатки и достоинства созданного текста и оценить его в сравнении с принятыми в данном дискурсивном сообществе нормами. Полученные результаты свидетельствуют о необходимости целенаправленной работы по совершенствованию навыков владения английским языком в целом и академическим английским для публикационных целей в частности. На следующем этапе планируется рассмотреть данные ресурсы с точки зрения их применимости для индивидуального пользования студентами и создать алгоритм работы над научным текстом с опорой на *AntConc* и *Coh-Metrix*. Данные по каждому конкретному критерию обработки текста могут составить основу теоретического и практического учебного материала для планируемого курса по написанию научной статьи на английском языке для студентов технического вуза.

Литература

1. Основы научной речи: Учеб. пособие для студ. нефилол. высш. учеб. заведений / Буре Н.А., Быстрых М.В., Вишнякова С.А. и др.; Под ред. Химика В.В., Волковой Л.Б. СПб.: СПбГУ; М.: Академия, 2003. 272 с.
2. *Добрынина О.Л.* Технология непрерывного иноязычного образования: учим студентов распознавать и исправлять ошибки в письменной речи // Непрерывное образование: XXI век. 2018. № 1 (21).
3. *Добрынина О.Л.* Грамматические ошибки в англоязычном академическом письме: причины появления и стратегии коррекции // Высшее образование в России. 2017. № 8/9 (215). С. 100–107.
4. *Добрынина О.Л.* Академическое письмо для публикационных целей: стилистические погрешности // Высшее образование в России. 2019. Т. 28. № 10. С. 38–49. DOI: https://doi.org/10.31992/0869-3617-2019-28-10-38-49
5. *Кузнецова Л.Б., Сучкова С.А.* Актив или пассив? «Я» или «Мы»? // Высшее образование в России. 2015. № 8-9. С. 143–148.
6. *Shpit E.I., Sobolevskaya O.V.* (2019) Analysing the level of academic writing literacy of TUSUR graduate students // Proc. of IEEE 2019 International multi-conference on engineering, computer and information sciences (SIBIRCON). Russia, Tomsk, Oct. 23–24, 2019. P. 0207–0211. URL: http://talgat.org/news/wp-content/uploads/2019/12/97.pdf

7. *Wallwork A.* English for Writing Research Papers. Springer Science+Business Media, LCC. 2011. 349 p.
8. *Лихачев Д.С.* Книга беспокойств. (Статьи, беседы, воспоминания). М.: Новости, 1991. 528 с.
9. *Laurence A.* AntConc: A Learner and Classroom Friendly, Multi-Platform Corpus Analysis Toolkit // IWLeL 2004: An Interactive Workshop on Language e-Learning, P. 7–13. URL: https://core.ac.uk/download/pdf/144458559.pdf
10. *Krajka J.* Corpora and Language Teachers: From Ready-Made to Teacher-Made Collections // CORELL: Computer Resources for Language Learning. 2007. No. 1. P. 36–55. URL: https://pdfs.semanticscholar.org/1b70/a6b768b1a1587442d28dd93685b5f3ad9cab.pdf
11. *Laurence A., Bowen M.* The language of mathematics: A corpus-based analysis of research article writing in a neglected field // Asian ESP J. 2013. Vol. 9. P. 5–25.
12. *Hidayat F.* Teaching Grammar by Induction to 21st Century Learners with Corpus Linguistics Technology // LIA International Conference and Cultural Events. Hyatt Regency Hotel, Yogyakarta, Indonesia. 2015. April 29 – May 1. URL: https://www.academia.edu/12166819/Teaching_Grammar_by_Induction_to_21st_Century_Learners_with_Corpus_Linguistics_Technology
13. *Рогачева В.Э.* Корпусный метод установления перевода терминологических единиц // Известия РГПУ им. А.И. Герцена. 2017. № 183. С. 101–107.
14. *McCarthy P., Lehenbauer B., Hall C., Duran N., Fujiwara Y., McNamara D.* A Coh-Metrix Analysis of Discourse Variation in the Texts of Japanese, American, and British Scientists. *Foreign Languages for Specific Purposes*. 2007. Vol. 6. URL: https://www.researchgate.net/publication/242322281_A_Coh-Metrix_Analysis_of_Discourse_Variation_in_the_Texts_of_Japanese_American_and_British_Scientists
15. *Duncan B., Hall C.* A Coh-Metrix Analysis of Variation among Biomedical Abstracts // Proceedings of the Twenty-Second International FLAIRS Conference. 2009. URL: https://www.researchgate.net/publication/221439065_A_Coh-Metrix_Analysis_of_Variation_among_Biomedical_Abstracts
16. *Duran N., Bellissens M., Taylor R., McNamara D.* Quantifying text difficulty with automated indices of cohesion and semantics // Proceedings of the 29th Annual Meeting of the Cognitive Science Society. 2007. P. 233–238.
17. *Dowell N., Graesser A., Cai Zh.* Language and Discourse Analysis with Coh-Metrix: Applications from Educational Material to Learning Environments at Scale // Journal of Learning Analytics. 2015. Vol. 3. DOI: https://doi.org/10.18608/jla.2016.33.5
18. *Солнышкина М.И., Кисельников А.С.* Параметры сложности экзаменационных текстов // Вестник ВолГУ. Серия 2: Языкознание. 2015. № 1. С. 99–107.
19. *McNamara D., McCarthy P., Graesser A., Cai Z.* Automated Evaluation of Text and Discourse with Coh-Metrix. Cambridge: Cambridge University Press, 2014. 278 p.
20. *McCarthy P., Jarvis S.* VOCD: A theoretical and empirical evaluation // Language Testing. 2007. Vol. 24. Issue 4. P. 459–488. DOI: https://doi.org/10.1177/0265532207080767
21. *Miller G., Beckwith R., Fellbaum C., Gross D., Miller K.J.* Introduction to WordNet: An on-line lexical database // Journal of Lexiography. 1990. Vol. 3. P. 235–244.
22. *Газизулина Л.Р.* Сложность и читабельность как критерии оценки учебного текста при обучении иностранному языку в неязыковом вузе // Мир науки, культуры и образования. 2019. № 1 (74). С. 372–374.

Благодарности. Работа выполнена при финансовой поддержке российского научного фонда (проект № 19-19-00424) в ТУСУРе.

Статья поступила в редакцию 01.04.20
Принята к публикации 13.06.20

Analysing Academic Texts with Computational Linguistics Tools

Elena I. Shpit – Senior language instructor, e-mail: forester_2007@mail.ru
Tomsk State University of Control Systems and Radioelectronics, Tomsk, Russia
Address: 40, Lenin Prospect, Tomsk, 634050, Russian Federation
Vassily N. Kurovskii – Dr. Sci. (Education), Prof., e-mail: v.kurovskii@yandex.ru
Tomsk State Pedagogical University, Tomsk, Russia
Address: 60, Kievskaya str., Tomsk, 634061, Russian Federation

Abstract. Writing academic texts in English introduces certain difficulties associated with translating Russian sentences with pronounced stylistic peculiarities, especially for young researchers who are just starting their publication activity. It seems impossible to study any genre without analysing examples of the discourse, which highlights the use of computational linguistics as it allows automating a lot of language and text processing mechanisms and generates relatively accurate quantitative results. The present study considers the application of *AntConc* and *Coh-Metrix* toolkits for analyzing master students' abstracts to research papers written for international English-language journals or conference proceedings (Learner Corpus) in comparison with international researchers' abstracts published in high-impact journals (Reference Corpus). The analysis conducted in the above-mentioned software tools revealed the drawbacks and strengths of master students' texts, allowed characterizing them on the words, sentence and discourse levels, as well as outlined the potentials of their use in teaching academic writing skills.

Keywords: academic writing, scientific text, computational linguistics, automated text analysis, *AntConc, Coh-Metrix*, specialized corpus

Cite as: Shpit, E.I., Kurovskii, V.N. (2020). Analysing Academic Texts with Computational Linguistics Tools. *Vysshee obrazovanie v Rossii = Higher Education in Russia*. Vol. 29, no. 7, pp. 89-103. (In Russ., abstract in Eng.)

DOI: https://doi.org/10.31992/0869-3617-2019-29-7-89-103

References

1. Bure, N.A., Bystrykh, M.V., Vishchnyakova, S.A., et al. (2003). *Osnovy nauchnoi rechi: uchebnoe posobie dlya studentov nefilol. vyssh. ucheb. zavedeniy* [Fundamentals of Scientific Register: Textbook for Non-Philology Tertiary Students]. Ed. by Khimik, V.V., Volkova, L.B. St. Peterburg: St. Petersburg State Univ. Moscow: Akademiya Publ. House. 272 p. (In Russ.)
2. Dobrynina, O.L. (2018). Technology of Lifelong Linguistic Education: Teaching Students to Recognize and Correct Errors in English Academic Writing. *Nepreryvnoe obrazovanie: XXI vek = Lifelong Education: The XXI century*. No. 1 (21). (In Russ., abstract in Eng.)
3. Dobrynina, O.L. (2017). Grammar Errors in Academic Writing in English: Causes and Strategies of Correction. *Vysshee obrazovanie v Rossii = Higher Education in Russia*. No. 8/9 (215), pp. 100-107. (In Russ., abstract in Eng.)
4. Dobrynina, O.L. (2019). Academic Writing for Publication Purposes: The Infelicities of Style. *Vysshee obrazovanie v Rossii = Higher Education in Russia*. Vol. 28, no. 10, pp. 38-49. DOI: https://doi.org/10.31992/0869-3617-2019-28-10-38-49. (In Russ., abstract in Eng.)
5. Kuznetsova, L.B., Suchkova, S.A. (2015). Active or Passive? "I" or "We"? *Vysshee obrazovanie v Rossii = Higher Education in Russia*. No. 8-9, pp. 143-148. (In Russ., abstract in Eng.)

6. Shpit, E.I., Sobolevskaya, O.V. (2019) Analysing the Level of Academic Writing Literacy of TUSUR Graduate Students. In: *Proc. of IEEE 2019 International multi-conference on engineering, computer and information sciences (SIBIRCON)*. Russia, Tomsk. 2019, Oct. 23–24, pp. 0207-0211. Available at: http://talgat.org/news/wp-content/uploads/2019/12/97.pdf
7. Wallwork, A. (2011). *English for Writing Research Papers*. Springer Science+Business Media, LCC, 349 p.
8. Likhachev, D.S. (1991). *Kniga Bespokoistv* [The Book of Worries]. Moscow: Novosti Publ., 528 p. (In Russ.)
9. Laurence, A. (2004). AntConc: A Learner and Classroom Friendly, Multi-Platform Corpus Analysis Toolkit. In: *IWLeL 2004: An Interactive Workshop on Language e-Learning*, pp. 7-13. URL: https://core.ac.uk/download/pdf/144458559.pdf
10. Krajka, J. (2007). Corpora and Language Teachers: From Ready-Made to Teacher-Made Collections. In: *CORELL: Computer Resources for Language Learning*, no. 1, pp. 36-55. Available at: https://pdfs.semanticscholar.org/1b70/a6b768b1a1587442d28dd93685b5f3ad9cab.pdf
11. Laurence, A., Bowen, M. (2013). The Language of Mathematics: A Corpus-Based Analysis of Research Article Writing in a Neglected Field. *Asian ESP J*. Vol. 9, pp. 5-25.
12. Hidayat, F. (2015). Teaching Grammar by Induction to 21st Century Learners with Corpus Linguistics Technology. In: *LIA International Conference and Cultural Events*. Hyatt Regency Hotel, Yogyakarta, Indonesia, April 29 – May 1. Available at: https://www.academia.edu/12166819/Teaching_Grammar_by_Induction_to_21st_Century_Learners_with_Corpus_Linguistics_Technology
13. Rogacheva, V. (2017). Corpus-Based Method of Terminology Translation. *Izvestiya RGPU imeni A.I. Gertsena = Izvestia: Herzen University Journal of Humanities & Sciences*. No. 183, pp. 101-107. (In Russ., abstract in Eng.)
14. McCarthy, P., Lehenbauer, B., Hall, C., Duran, N., Fujiwara, Y., McNamara, D. (2007). A Coh-Metrix Analysis of Discourse Variation in the Texts of Japanese, American, and British Scientists. *Foreign Languages for Specific Purposes*. Vol. 6. Available at: https://www.researchgate.net/publication/242322281_A_Coh-Metrix_Analysis_of_Discourse_Variation_in_the_Texts_of_Japanese_American_and_British_Scientists
15. Duncan, B., Hall, C. (2009). A Coh-Metrix Analysis of Variation among Biomedical Abstracts. In: *Proceedings of the Twenty-Second International FLAIRS Conference*. Available at: https://www.researchgate.net/publication/221439065_A_Coh-Metrix_Analysis_of_Variation_among_Biomedical_Abstracts
16. Duran, N., Bellissens, M., Taylor, R., McNamara, D. (2007). Quantifying Text Difficulty with Automated Indices of Cohesion and Semantics. In: *Proceedings of the 29th Annual Meeting of the Cognitive Science Society*, pp. 233-238.
17. Dowell, N., Graesser, A., Cai, Zh. (2015). Language and Discourse Analysis with Coh-Metrix: Applications from Educational Material to Learning Environments at Scale. *Journal of Learning Analytics*. No. 3. DOI: https://doi.org/10.18608/jla.2016.33.5
18. Solnyshkina, M.I., Kiselnikov, A.S. (2015). The Indices of Examination Texts Complexity. *Vestnik Volgogradskogo gosudarstvennogo universiteta. Seriya 2. Jazykoznanije = Science Journal of VolSU. Linguistics*. No. 1, pp. 99-107. (In Russ., abstract in Eng.)
19. McNamara, D., McCarthy, P., Graesser, A., Cai, Zh. (2014). Automated Evaluation of Text and Discourse with Coh-Metrix. Cambridge: Cambridge University Press, 278 p.
20. McCarthy, P., Jarvis, S. (2007). VOCD: A Theoretical and Empirical Evaluation. *Language Testing*, Vol. 24, Issue 4, pp. 459-488. DOI: https://doi.org/10.1177/0265532207080767

21. Miller, G., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.J. (1990). Introduction to WordNet: An on-line lexical database. *Journal of Lexiography*, Vol. 3, pp. 235-244.
22. Gazizulina, L.R. (2019). Complexity and Readability Criteria for the Assessment of Academic Text in Foreign Language Training in a Non-Language Higher Education Institution. *Mir nauki, kulturi i obrazovania = The World of Science, Culture and Education*. No. 1 (74), pp. 372–374. (In Russ., abstract in Eng.)

Acknowledgement. The reported study was funded by Russian Science Foundation (project № 19-19-00424) in TUSUR.

The paper was submitted 01.04.19
Accepted for publication 13.06.20