УДК 81'322.2

Грачева Д.А.
студент
Фролова А.С.
студент
Черкасова В.М.
студент
Ульяновский государственный технический университет
Россия, Ульяновск

ТЕХНОЛОГИЯ СОЗДАНИЯ СПЕЦИАЛИЗИРОВАННОГО ДВУЯЗЫЧНОГО КОРПУСА ТЕКСТОВ В ОБЛАСТИ AGILE-МЕТОДОЛОГИЙ

Аннотация: *Актуальность исследования обусловлена возрастающей потребностью в специализированных корпусах текстов для изучения профессиональных дискурсов, в частности в rapidly развивающейся области управления проектами на основе гибких методологий. Цель работы — описать и апробировать методику создания двуязычного (англо-русского) корпуса текстов по Agile-тематике с использованием доступного инструментария. В статье применяются методы корпусной лингвистики: сбор и фильтрация текстовых данных, автоматическая обработка с помощью конкордансера AntConc, частотный анализ. Результатом исследования является сформированный сопоставимый корпус (9 англоязычных и 9 русскоязычных текстов) и апробированная пошаговая технология его создания. Практическая значимость работы заключается в возможности тиражирования предложенной методики для построения корпусов в других предметных областях.*

Ключевые слова: *корпусная лингвистика, специализированный корпус, AntConc, Agile, управление проектами, частотный анализ, сопоставимый корпус.*

Gracheva D.A.
Student
Frolova A.S.
Student
Cherkasova V.M.
Student
Ulyanovsk State Technical University
Russia, Ulyanovsk

TECHNOLOGY FOR CREATING A SPECIALIZED BILINGUAL TEXT CORPUS IN THE FIELD OF AGILE METHODOLOGIES

Abstract*: The relevance of the study is determined by the growing need for specialized text corpora for the study of professional discourses, particularly in the rapidly developing field of project management based on agile methodologies. The aim of the work is to describe and test a methodology for creating a bilingual (English-Russian) corpus of texts on Agile topics using accessible tools. The article employs corpus linguistics methods: collection and filtering of text data, automatic processing using the AntConc concordancer, and frequency analysis. The result of the study is a formed comparable corpus (9 English-language and 9 Russian-language texts) and an approved step-by-step technology for its creation. The practical significance of the work lies in the possibility of replicating the proposed methodology for building corpora in other subject areas.*

Keywords*: corpus linguistics, specialized corpus, AntConc, Agile, project management, frequency analysis, comparable corpus.*

Введение

Современная корпусная лингвистика предоставляет исследователю широкий спектр методов и инструментов для изучения языковых явлений на эмпирическом материале. В.П. Захаров и С.Ю. Богданова подчёркивают, что современный этап развития корпусной лингвистики характеризуется доминированием автоматизированных методов, однако полная автоматизация остаётся недостижимой: задачи отбора источников, разработки системы разметки и верификации результатов по-прежнему требуют вмешательства лингвиста-эксперта [1]. О.В. Нагель отмечает, что корпусный подход оптимален для наглядного представления таких аспектов языка, как историческая, географическая и социальная вариация, а также изменения в языковой системе [2].

Особую актуальность приобретает разработка специализированных корпусов, отражающих конкретные профессиональные дискурсы. Гибкие методологии управления проектами (Agile, Scrum, Kanban) представляют собой динамично развивающуюся область, лексический состав которой требует систематического изучения на материале как английского, так и русского языков. Несмотря на значительное количество работ, посвящённых общим принципам корпусной лингвистики, конкретные технологии создания узкоспециализированных двуязычных корпусов в области IT-дискурса описаны недостаточно полно, что обусловливает необходимость настоящего исследования.

Цель работы — описать и апробировать технологию создания двуязычного сопоставимого корпуса текстов в предметной области Agile и управления проектами. Достижение цели предполагает решение следующих задач: (1) обосновать критерии отбора текстового материала; (2) описать процедуру предварительной обработки текстов, включая нормализацию и фильтрацию; (3) обосновать выбор программного инструмента AntConc и продемонстрировать его возможности для анализа

сформированного корпуса; (4) верифицировать репрезентативность полученного корпуса.

Гипотеза исследования состоит в том, что применение поэтапной методологии сбора, обработки и анализа текстовых данных с использованием свободно распространяемого инструментария позволяет создать репрезентативный специализированный корпус, пригодный для решения лингвистических задач.

Материалы исследования

Эмпирическую базу исследования составили два подкорпуса текстов, объединённых общей тематикой — гибкие методологии управления проектами (Agile, Scrum, Kanban) и их развитие в условиях цифровой трансформации. Все тексты относятся к публицистическому и экспертно-аналитическому дискурсу и адресованы профессиональной аудитории (IT-специалистам, проектным менеджерам, руководителям).

Англоязычный подкорпус включает 9 статей, опубликованных на отраслевых платформах и в профессиональных блогах: ITNEXT.io, Medium / Product Coalition, iPeople.nl, Age of Product, Forrester, StarAgile, RMC Learning Solutions, Medium / Reader's Club, Medium / Michael Head. Общий объём — 7 449 слов, 41 924 знака без пробела.

Русскоязычный подкорпус включает 9 статей, опубликованных на платформе Habr.com — одном из крупнейших русскоязычных сообществ IT-специалистов. Тематика статей охватывает проблемы внедрения Agile в российских компаниях, эволюцию Scrum, вопросы масштабирования гибких методологий. Общий объём — 7 723 слова, 58 503 знака без пробела.

Обеспечение репрезентативности и сбалансированности корпуса

При формировании корпуса соблюдались следующие принципы репрезентативности. Во-первых, принцип тематической однородности: все тексты отбирались по критерию принадлежности к дискурсу об Agile и

управлении проектами, что обеспечивает внутреннюю целостность корпуса. Во-вторых, принцип жанровой однородности: использовались только публицистические и экспертно-аналитические статьи, исключающие техническую документацию, маркетинговые материалы и академические тексты. В-третьих, принцип сопоставимости объёмов: разница в объёме подкорпусов составляет менее 4%, что допустимо для сопоставительного анализа. В-четвёртых, принцип хронологической релевантности: все тексты опубликованы в течение последних двух лет, что отражает актуальное состояние дискурса.

Обоснование выбора инструментария

Для автоматической обработки использовался конкордансер AntConc, разработанный Л. Энтони [3]. Выбор данного инструмента обусловлен несколькими факторами. Во-первых, программа является свободно распространяемой и кроссплатформенной, что делает её доступной для исследователей независимо от технических ресурсов. Во-вторых, AntConc поддерживает работу с текстами в кодировке UTF-8, что критически важно для обработки русскоязычных текстов. В-третьих, программа предоставляет полный набор базовых функций корпусного анализа, необходимых для решения поставленных задач.

Л. Энтони отмечает, что функциональность программных инструментов во многом определяет доступные исследователю методы корпусного анализа [3]. AntConc включает семь основных инструментов: Word List (построение частотных списков с сортировкой по частоте, алфавиту или окончаниям), Keyword List (выявление статистически значимых ключевых слов), Concordance (конкорданс в формате KWIC), Concordance Plot (визуализация распределения вхождений), Clusters / N-Grams (автоматическое извлечение устойчивых последовательностей), Collocates (вычисление статистически значимых коллокаций), File View (просмотр полного контекста в исходном файле). Важной особенностью

программы является поддержка регулярных выражений, что расширяет возможности поиска.

В отличие от коммерческих платформ (например, Sketch Engine), AntConc не имеет встроенного морфологического анализатора, однако для задач частотного и контекстуального анализа на материале текстов с ограниченной морфологической вариативностью (преимущественно терминологическая лексика) данное ограничение не является критическим.

Методология обработки

Обработка текстов проводилась в несколько последовательных этапов.

На первом этапе производилась предварительная очистка текстов от шумовой информации: удалялись метаданные (заголовки HTML-страниц, даты публикации, имена авторов, навигационные элементы), рекламные вставки, ссылки на источники в тексте. Для каждого текста сохранялась только содержательная часть.

На втором этапе осуществлялась нормализация текстов: приведение к единой кодировке (UTF-8), удаление лишних пробелов и символов форматирования, унификация кавычек и дефисов. Для русскоязычных текстов дополнительно проводилась проверка на наличие опечаток.

На третьем этапе в программе AntConc строились частотные списки для каждого подкорпуса. Полученные списки экспортировались в табличный формат для дальнейшей обработки.

На четвёртом этапе из частотных списков удалялись стоп-слова — лексические единицы, не несущие терминологической ценности для исследуемой предметной области. В работе применялся комплексный подход к фильтрации, основанный на сочетании лингвистических критериев и контекстуальной релевантности. К стоп-словам отнесены: (а) общеупотребительная и абстрактная лексика (работа, процесс, изменение, задача; work, process, change, task); (б) лексика традиционного каскадного управления, не являющаяся неотъемлемой частью Agile (контракт,

бюджет, отчёт; contract, budget, report); (в) общенаучная и оценочная лексика (базовый, сложный, простой; basic, complex, simple); (г) глаголы широкой семантики и модальные конструкции (делать, использовать, помогать; do, use, help, can, must); (д) служебные части речи (предлоги, союзы, частицы, местоимения).

На пятом этапе формировались итоговые репрезентативные списки тематической лексики (по 100 терминов для каждого языка), включающие только единицы, непосредственно относящиеся к фреймворкам Agile (Scrum, Kanban, XP, Lean), ролям (product owner, scrum master), артефактам (backlog, sprint, increment), событиям (retrospective, daily stand-up) и ключевым принципам (iterative, continuous improvement, transparency).

Результаты исследования

В результате применения описанной технологии был сформирован двуязычный специализированный корпус текстов, содержащий 18 текстов (9 на английском и 9 на русском языке) общей тематики. Корпус сохранён в виде структурированного набора файлов в формате .txt (по одному файлу на каждый текст), что обеспечивает совместимость с программным обеспечением для корпусного анализа.

Построены частотные списки для каждого подкорпуса, выделены высокочастотная (от 190 до 10 употреблений для английского, от 92 до 10 для русского), среднечастотная (9–6 употреблений) и низкочастотная (5–1 употребление) зоны. Составлены репрезентативные списки по 100 тематических терминов для каждого языка.

Анализ высокочастотной лексики английского подкорпуса показал доминирование терминов agile (190 употреблений), scrum (95), team (69), organization (37), practice (29), development (27), framework (24), ai (23), learning (23). В русскоязычном подкорпусе наиболее частотными оказались: команда, проект, разработка, процесс, работа, продукт, управление, отдел, документация, контракт. Данное распределение

подтверждает тематическую однородность корпуса и его соответствие заявленной предметной области.

Программа AntConc продемонстрировала эффективность при решении задач конкордансного анализа. Вкладка Concordance позволила проследить контексты употребления ключевых терминов, вкладка N-Grams — выделить устойчивые словосочетания (например, product owner, sprint backlog, daily standup в английском подкорпусе; канбан-доска, scrum-мастер, спринт-планирование в русском), вкладка Collocates — выявить статистически значимые сочетания. Функция Concordance Plot позволила визуализировать распределение вхождений терминов по текстам и оценить равномерность их употребления.

Для верификации репрезентативности корпуса построены графики Ципфа для обоих подкорпусов. Полученные распределения соответствуют типичному для естественных языков степенному закону: небольшое количество высокочастотных единиц покрывает значительную долю текста, в то время как основная масса лексики представлена единичными вхождениями. Данный результат подтверждает естественность и репрезентативность сформированного корпуса.

Описанная технология создания специализированного корпуса имеет ряд преимуществ. Во-первых, она не требует специализированного программного обеспечения, выходящего за рамки свободно распространяемых инструментов. Во-вторых, поэтапный характер методологии позволяет контролировать качество на каждом этапе. В-третьих, воспроизводимость технологии обеспечивает возможность её применения другими исследователями.

Вместе с тем следует отметить ограничения метода. Отсутствие встроенного лемматизатора в AntConc приводит к тому, что разные словоформы одного слова учитываются как отдельные единицы, что может несколько искажать частотные показатели. Для русского языка данная проблема более актуальна в силу развитой морфологии. В

настоящем исследовании этот недостаток частично компенсировался тем, что терминологическая лексика (составляющая ядро корпуса) характеризуется ограниченной морфологической вариативностью.

Ещё одним ограничением является ручной характер этапа фильтрации стоп-слов. Хотя данная процедура обеспечивает высокую точность, она трудоёмка и зависит от компетентности исследователя в предметной области. Автоматизация данного этапа с использованием готовых списков стоп-слов и алгоритмов машинного обучения представляет собой перспективное направление дальнейшей работы.

Заключение

Данная технология создания специализированного двуязычного корпуса текстов доказала свою практическую применимость. Использование свободно распространяемого инструмента AntConc позволило решить полный цикл задач — от построения частотных списков до контекстуального анализа. Сформированный корпус объёмом 18 текстов (15 172 слова суммарно) является репрезентативным для изучаемой предметной области, что подтверждено анализом распределения частот (графики Ципфа) и тематической однородностью высокочастотной лексики.

Как отмечает В.П. Захаров, в реальной практике доминирует смешанный подход, при котором автоматическая обработка сочетается с последующей верификацией и ручной корректировкой [1], что полностью подтвердилось в ходе настоящего исследования. Данный подход оказался особенно эффективным при работе с русскоязычными текстами, где автоматическая обработка осложняется морфологическими особенностями языка.

Методика может быть масштабирована при создании корпусов в других профессиональных дискурсах. В качестве перспектив дальнейшей работы рассматриваются: расширение корпуса за счёт включения текстов других жанров (техническая документация, стенограммы конференций,

интервью); автоматизация этапа фильтрации стоп-слов; применение методов лемматизации для повышения точности частотного анализа; использование сформированного корпуса для решения конкретных лингвистических задач (анализ коллокаций, выявление терминологических кластеров, сопоставительный анализ дискурсов).

Использованные источники:

1. Захаров В.П., Богданова С.Ю. Корпусная лингвистика: учебник. Иркутск: ИГЛУ, 2011. 161 с.

2. Нагель О.В. Корпусная лингвистика и ее использование в компьютеризированном языковом обучении // Вестник Томского государственного университета. 2008. № 2(16). С. 53–59.

3. Anthony L. A critical look at software tools in corpus linguistics // Linguistic Research. 2013. Vol. 30, No. 2. P. 141–161.