Научная статья
УДК 81'23
DOI 10.52070/2542-2197_2023_9_877_35

Моделирование порождения текста на основе данных кейлоггеров и дистрибутивных семантических моделей: обоснование методологии и программа исследований

Т. А. Литвинова

Воронежский государственный педагогический университет, Воронеж, Россия
centr_rus_yaz@mail.ru

Аннотация. В современной экспериментальной психолингвистике активно используются кейлоггеры – программы для регистрации клавиатурного почерка. Их применение позволило получить новые научные данные о процессе порождения текста, но в то же время привело к появлению разного рода противоречивых сведений. В статье предлагается дополнить методологию исследований порождения текста с использованием кейлоггеров данными дистрибутивных семантических моделей о семантическом расстоянии между словами.

***Ключевые слова*:** порождение текста, кейлоггер, экспериментальная психолингвистика, дистрибутивные семантические модели, минимальные единицы порождения текста, клавиатурный текст

Благодарность. Исследование выполняется в Воронежском государственном педагогическом университете при поддержке гранта Российского научного фонда № 21-78-10148.

***Для цитирования*:** Литвинова Т. А. Моделирование порождения текста на основе данных кейлоггеров и дистрибутивных семантических моделей: обоснование методологии и программа исследований // Вестник Московского государственного лингвистического университета. Гуманитарные науки. 2023. Вып. 9 (877). С. 35–40. DOI 10.52070/2542-2197_2023_9_877_35

Original article

Modelling Writing Based on Keyloggers Data and Distributional Semantic Models: Justification of Methodology and Research Program

Tatiana A. Litvinova

Voronezh State Pedagogical University, Voronezh, Russia
centr_rus_yaz@mail.ru

Abstract. In modern experimental psycholinguistics, keyloggers (i. e., programs for recording keyboard behavior) are actively used. Their application made it possible to obtain new scientific data on the process of text production, but at the same time led to the emergence of various contradictory information. We propose to supplement the methodology of text generation studies using keyloggers with data from distributional semantic models on the semantic distance between words.

***Keywords*:** writing process, keylogger, experimental psycholinguistics, distributional semantic models, minimal units of text production, keyboard text

Acknowledgment. The study is supported by Russian Science Foundation (grant N 21-78-10148).

***For citation*:** Litvinova, T. A. (2023). Modelling Writing Based on Keyloggers Data and Distributional Semantic Models: Justification of Methodology and Research Program. Vestnik of Moscow State Linguistic University. Humanities, 9(877), 35–40 10.52070/2542-2197_2023_9_877_35

ВВЕДЕНИЕ

В большинстве направлений лингвистики текст традиционно исследуется как законченное речевое произведение, продукт речевой деятельности. Однако для решения ряда фундаментальных проблем языкознания, связанных в частности с построением и развитием теорий порождения текста, критически важным является анализ процесса порождения речевого высказывания. Появление новых программных средств, в частности регистраторов нажатий клавиш (кейлоггеров), обеспечило исследователя информацией о длительности нажатий на клавиши и пауз, вследствие чего оказало значительное влияние на развитие работ в области порождения текста.

Кейлоггеры позволяют детально анализировать процесс текстопорождения в различных условиях коммуникации. Причем новые программные средства в отличие, например, от айтрекеров не оказывают воздействия на респондентов.

Применение подобных программных средств, позволяющих реконструировать процесс создания текста, стало по сути «золотым стандартом» в области исследований письма (writing research). В большинстве работ, относящихся к означенной области лингвистики, тексты анализируются с использованием программы InputLog [Leijten, Waes, 2013], лингвистический модуль которой поддерживает голландский и английский языки.

Русский язык находится на периферии подобных исследований, в то время как отечественные ученые внесли важнейший вклад в развитие психофизиологии письма, разработку моделей порождения речи (труды Л. С. Выготского, А. Р. Лурия, А. А. Леонтьева, Т. В. Ахутиной, Н. И. Жинкина, И. А. Зимней и т. д.). Как указывает Т. В. Ахутина, научные взгляды А. А. Леонтьева во многом определили многие современные исследования не только отечественных, но и ряда зарубежных ученых по данной проблеме (см. подробнее об этом: [Глухов, 2005]).

Однако в области исследований письма с использованием записей процесса текстопорождения существует много нерешенных и малоизученных проблем. В частности, многие исследователи отмечают, что лингвистические аспекты процесса порождения текста мало изучены: ученые уделяют внимание преимущественно характеру и местоположению пауз, а также анализу самоисправлений (ревизий). Как отмечается в работе [Mahlow, Ulasik, Tuggener, 2022], попыток глубокого лингвистического анализа процесса письма до сих пор предпринято не было, несмотря на то, что необходимость такого анализа неоднократно подчеркивалась.

В настоящей работе обоснована необходимость расширения методологии исследования порождения клавиатурного текста. Методологические новации, в свою очередь, обеспечиваются моделированием данных о семантическом расстоянии между лексемами. Эти данные могут быть успешно извлечены из дистрибутивных семантических моделей.

В данной работе предлагаются не только отдельно взятые методологические новации, но также излагается конкретная исследовательская программа по работе с ними.

ИСПОЛЬЗОВАНИЕ ДИСТРИБУТИВНО-СЕМАНТИЧЕСКИХ МОДЕЛЕЙ И КЕЙЛОГГЕРОВ В СОВРЕМЕННЫХ МОДЕЛЯХ ПОРОЖДЕНИЯ ТЕКСТА

Современные модели порождения текста, разработанные зарубежными исследователями, сконцентрированы на объяснении когнитивных процессов и воспроизводят язык как средство передачи смысла (подробнее об этом: [Mahlow, Ulasik, Tuggener, 2022]).

Как отмечалось выше, отечественные психолингвисты разработали ряд моделей порождения связного высказывания, в которых уделялось внимание различным аспектам текстопорождения, в том числе лингвистическим. Отдельную научную проблему, которую решают отечественные исследователи, составляет выявление минимальных структурных единиц текстопорождения (см., например, работу [Кибрик, Подлесская, 2009], в которой рассматривается элементарная дискурсивная единица (ЭДЕ) — «квант устного дискурса, минимальный шаг, при помощи которого говорящий продвигает дискурс вперед»). Однако, насколько нам известно, подобных исследований на материале письменных текстов для русского языка не существует.

Феномен текстопорождения мало исследован и в мировой науке в целом. В качестве одного из единичных примеров описания лингвистической природы аналога ЭДЕ в письменных текстах – bursts – возможно привести монографию французских авторов [Cislaru, Olive, 2018]. Однако в данной работе используется единственно паузальный критерий выделения элементарных единиц. Никакие другие критерии не рассматриваются.

Как известно, среди исследователей не достигнуто консенсуса в вопросе о том, какое пороговое значение пауз отделяет «когнитивные паузы» (то есть паузы, отражающие процесс построения высказывания, развертывание смысла) от «некогнитивных» [Baaijen, Galbraith, de Glopper, 2012]. Очевидно, однако, то, что различные авторы обладают

различной скоростью письма. Кроме того, в процессе создания текста авторы могут отвлекаться на посторонние мысли и т. д. Поэтому исследователи предпринимают попытки выделить индивидуальные пороговые значения на основании медианной длительности межклавишных интервалов и т. д. (Подробнее об этом см. в работе [Baaijen, Galbraith, de Glopper, 2012]). Очевидно, что опора только на паузальный критерий, без оценки контекста и без дополнительной лингвистической информации не может служить основой для получения достоверных результатов моделирования процесса порождения печатного текста.

Ряд исследовательских групп предпринимают попытки обогатить данные кейлоггеров лингвистической информацией путем применения средств автоматической лингвистической разметки, а именно частеречной и частичной синтаксической разметки (вследствие низкой точности). В работах последних лет представлены датасеты, содержащие данные кейлоггеров и размеченные по типам ревизий – как вручную, так и автоматически [Conijn, 2022], а также аннотацию по различным элементам текстопорождения [Miletić, 2022]. Следует также отметить, что перевод «сырых» данных кейлоггеров в лингвистически значимый формат (то есть перевод информации на уровень слов и пауз между ними) представляет собой отдельную техническую задачу. Она решается по-разному, что также влияет на результаты исследований.

Предпринимаются попытки визуализации данных записи процесса письма. Однако о механизмах визуализации говорится только применительно к символам и словам [Goodkind, 2021].

Появление кейлоггеров, специально предназначенных для академических исследований, привело к активизации исследований процесса создания письменного текста. В подобных работах не только уточняются теоретические концепции текстопорождения, но также решаются практико-ориентированные задачи. В частности, отдельное направление исследований связано с изучением влияния стратегий порождения текста (выделяемых на основе числа, продолжительности пауз и их местоположения) на его качество (на материале текстов как на родном, так и на иностранном языках). Исследуется эффективность рекомендаций по улучшению текста на основе данных кейлоггера [Vandermeulen, Steendam, Rijlaarsdam, 2023]. Многочисленными научными группами изучаются особенности порождения текста лицами с нейрогенеративными заболеваниями [Meulemans et al., 2022], перенесшими инсульт и т. д.

Если бы анализ текстопорождения, основанный преимущественно на исследовании пауз, был бы обогащен лингвистической информацией, возникла бы возможность разработки лечебных программ и систем социальной психотерапии. К ним относится диагностика нейрогенеративных заболеваний, система рекомендаций по улучшению построения связного текста и др. Поэтому исследователи предпринимают попытки лингвистической разметки данных кейлоггеров. Однако, как отмечалось выше, в основном работа, направленная на выявление механизмов текстопорождения, ограничивается частеречной разметкой. Особую ценность представляет ручная разметка, хотя ее выполнение трудоемко.

СЕМАНТИЧЕСКАЯ СВЯЗНОСТЬ ТЕКСТА

Представляется продуктивным обогащение данных кейлоггеров (после их приведения к пригодному для анализа виду) информацией о семантической близости слов и их последовательностей. Данная информация выявляется на основе векторных семантических моделей разных видов (как контекстно-независимых – word2vec, GloVe и др., так и контекстно-зависимых, например, BERT).

Исследование семантической связности текстов на основе подобных моделей является актуальным направлением исследований в медицине, в частности в неврологии и психиатрии (см. подробнее об этом: [Holmlund et al., 2022]). В подобных работах анализ текст рассматривается как диагностический инструмент и используется для выявления тех или иных заболеваний на ранней стадии, для описания динамики состояния пациента. В означенных случаях текст выступает как законченный продукт речевой деятельности. О работах, где использовались бы одновременно данные кейлоггера и семантической близости, нам неизвестно. Между тем текстовая диагностика сегодня получает все большее распространение.

Широко используется подобная методология в исследованиях креативности [Heinen, Johnson, 2018].

Создаются специальные веб-ресурсы и программные средства для расчета семантической связности текстов (см., например, [Beaty, Johnson, 2021]. Отметим, что названная работа, вышедшая менее чем два года назад, имеет более 100 цитирований, что свидетельствует о большом интересе научного сообщества к данной проблеме.

Однако подобные исследования сопряжены с проблемой выбора метрики семантической связности (какую меру брать для анализа – среднее семантическое расстояние между словами в тексте, минимальное расстояние, стандартное отклонение и т. д.; какую ширину окна выбрать и т. д.). Указанные вопросы по-разному решаются исследователями,

вследствие чего сопоставить полученные результаты может быть затруднительно.

В работе [Holmlund at al, 2022] впервые отмечается необходимость учета темпоральной информации как возможного решения названных проблем, но применительно к устному тексту. Работ по анализу письменного текста, направленных на исследование семантического расстояния между различными его элементами с учетом данных о паузах между ними, насколько нам известно, нет.

ПРЕДЛАГАЕМОЕ РЕШЕНИЕ: ОБЪЕДИНЕНИЕ ПОДХОДОВ

Приведенный анализ показывает, что текущий уровень развития исследований в области *writing research*, с одной стороны, и семантической связности текста – с другой, требует симбиоза методологий названных научных направлений, что будет способствовать развитию каждого из них.

Мы впервые предлагае впервые использовать для моделирования процесса письма, наряду с данными кейлоггера, данные о семантической близости единиц текстопорождения (слов, последовательностей из *n*-слов), вычисленные на основе векторных семантических моделей.

Полагаем, что моделирование процесса порождения письменного (печатного) текста путем сопоставления данных кейлоггера о паузах между различными последовательными элементами процесса текстопорождения и преодобученных семантических моделей о семантическом расстоянии между названными элементами в динамике (по ходу развертывания смысла связного речевого высказывания) необходимо проводить с учетом личностных характеристик авторов текстов и типа коммуникативного задания. Неотъемлемым элементом программы исследований является интерактивная визуализация результатов выполненного моделирования.

Исследование связи между значением семантического расстояния и длительностью пауз для элементов разной длины (мы рассматриваем в текущем проекте линейную структуру письменного текста) позволит расширить представление о лингвистических характеристиках процесса создания текста, в частности, о минимальных структурных единицах порождения письменного (печатного) текста.

Предлагаемая нами программа исследований предположительно может включать в себя следующие этапы:

1) **создание датасета нового типа**:
   - разработка анкеты, включающей в себя опросники, направленные на выявление особенностей когнитивной сферы респондентов, описание условий моделируемой коммуникативной ситуации (для текстов, созданных по заданию экспериментатора);
   - рекрутинг респондентов, сбор экспериментальных данных (текстов, созданных по заданию экспериментатора) и текстов, созданных в условиях естественной коммуникации, с использованием различных средств регистрации нажатий клавиш, предназначенных для академических исследований;
   - разработка скрипта по переводу «сырых» данных кейлоггера в данные, пригодные для дальнейшего анализа (слова, последовательности слов и длительности пауз между ними);
2) **экспериментальные исследования:**
   - расчет индивидуальных паузальных критериев, классификация их по длительности;
   - выделение единиц анализа – последовательностей слов разной длины на основании индивидуальных паузальных критериев (линейные события текстопорождения);
   - расчет семантической близости между выделенными на предыдущем этапе единицами с использованием различных семантических моделей, а также с использованием агрегированного критерия, расчет значений семантического расстояния между словами на границах пауз);
   - установление отношений между семантическим расстоянием и длительностью пауз с учетом индивидуального фактора, характеристик авторов и условий коммуникации;
   - выявление, на основании комбинированного паузально-семантического критерия, минимальных структурных единиц текстопорождения и определение их лингвистической природы (ручная разметка);
   - анализ динамики показателей семантической связности с учетом индивидуального фактора, характеристик авторов и условий коммуникации;
3) **визуализация результатов моделирования:**
   - разработка веб-приложения для интерактивной визуализации динамики семантического расстояния между выделенными структурными единицами на основании паузально-семантического критерия.

В лаборатории корпусной идиолектологии ВГПУ осуществлены наработки по названному направлению исследований.

В частности, сформирована база исследований в области экспериментальной идиолектной психолингвистики – первый русскоязычный датасет, содержащий видеозаписи процесса порождения текста, а также устные тексты информантов.

Датасет содержит как устные, так и письменные монологические и диалогические тексты респондентов, а также разметку по элементарным дискурсивным единицам (для устных текстов), разметку в программе ELAN – для видеозаписей письменных текстов.

На сформированном датасете проведены пилотные исследования клавиатурно-опосредованной индивидуальной речевой деятельности носителя русского языка в идентификационном аспекте, экспериментально и теоретически обосновано использование в идентификационных идиолектных исследованиях новых динамических стилеметрических маркеров, которые отражают особенности протекания идиолектной деятельности в реальном времени и взаимодействия субъекта с техническим устройством (более подробно об этом см.: [Litvinova, 2020]).

ЗАКЛЮЧЕНИЕ

Таким образом, нами предложена методология, комбинирующая подходы, связанные с новейшими разработками в области экспериментальной и компьютерной лингвистики, – с исследованием клавиатурного почерка и дистрибутивно-семантическими моделями.

Описанная методология позволит получить следующие результаты:

- первый датасет на русскоязычном материале, содержащий данные о процессе порождения текста, финальный продукт (тексты) и многослойную разметку;
- новые научные данные об отношениях между семантическим расстоянием и длительностью пауз с учетом индивидуального фактора, характеристик авторов и условий коммуникации;
- данные о лингвистических характеристиках и типах минимальных структурных единиц текстопорождения, выделенных на основании комбинированного паузально-семантического критерия;
- новые научные данные о динамике показателей семантической связности с учетом индивидуального фактора, характеристик авторов и условий коммуникации;
- веб-приложение для интерактивной визуализации динамики процесса порождения текста на основе данных кейлоггера и векторных семантических моделей.

Ожидаемые научные результаты не только позволят расширить теоретические представления о процессе порождения письменного текста, его этапах, структурных единицах, влиянии на названный процесс характеристик авторов и коммуникативной ситуации, но также будут способствовать решению различных прикладных задач. В частности, результаты анализа текстопорождения могут быть использованы:

- для разработки методов оценки динамики состояния пациентов с различными нейрогенеративными заболеваниями;
- для разработки методов диагностирования креативности на основе анализа текста;
- для повышения эффективности рекомендаций по улучшению качества текста;
- при разработке программных продуктов, направленных на улучшение систем коррекции ввода текста.

СПИСОК ИСТОЧНИКОВ

1. Leijten M., Van Waes L. Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes // Written Communication. 2013. № 30(3). P. 358–392.
2. Глухов В.П. Основы психолингвистики. М.: АСТ: Астрель, 2005.
3. Mahlow C., Ulasik M.A. Tuggener D. Extraction of transforming sequences and sentence histories from writing process data: a first step towards linguistic modeling of writing // Read Writ. 2022.
4. Рассказы о сновидениях: Корпусное исследование устного русского дискурса / Кибрик А. А., Подлесская В. И. (ред.) М.: ЯСК, 2009.
5. Cislaru G., Olive T. Le processus de textualisation. Analyse des unités linguistiques de performance écrite = Анализ порождения текста. Минимальные единица порождения текста. Louvain-la-Neuve: De Boeck Supérieur, 2018.
6. Baaijen V. M., Galbraith D., de Glopper, K. Keystroke analysis: Reflections on procedures and measures // Written Communication. 2012. № 29(3). P. 246–277.
7. Conijn R. et al. A product-and process-oriented tagset for revisions in writing // Written Communication. 2022. T. 39, № 1. P. 97–128.
8. Miletic A. et al. Pro-TEXT: an Annotated Corpus of Keystroke Logs // Proceedings of LREC. 2022. P. 1732–1739
9. Goodkind A. TypeShift: A User Interface for Visualizing the Typing Production Process. 2021. arXiv:2103.04222.
10. Vandermeulen N., Steendam E. Van, Rijlaarsdam G. Writing Process Feedback Based on Keystroke Logging and Comparison with Exemplars: Effects on the Quality and Process of Synthesis Texts // Written Communication. 2023. № 40 (1). P. 90–144.

11. Meulemans, C. et al. (2022). Cognitive Writing Process Characteristics in Alzheimer's Disease // Frontiers in psychology. № 13. DOI: 10.3389/fpsyg.2022.872280.
12. Holmlund T. B. et al. Towards a temporospatial framework for measurements of disorganization in speech using semantic vectors. Schizophrenia research. 2022. DOI: 10.1016/j.schres.2022.09.020.
13. Heinen D.J.P., Johnson D. R. Semantic distance: An automated measure of creativity that is novel and appropriate // Psychology of Aesthetics, Creativity, and the Arts. 2018. № 12 (2). P. 144.
14. Beaty R. E., Johnson D. R. Automating creativity assessment with SemDis: An open platform for computing semantic distance // Behavior research methods. 2021. 53(2). P. 757–780.
15. Litvinova T. Process-oriented characteristics of an idiolect for authorship attribution of heterogeneous texts: A pilot study // CEUR Workshop Proceedings. 2020. Vol. 2780. P. 3–6.

REFERENCES

1. Leijten, M., Van Waes, L. (2013). Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. Written Communication, 30(3), 358–392.
2. Gluxov, V. P. (2005). Osnovy psyholingvistiki = Foundations of psycholinguistics. Moscow: AST: Astrel. (In Russ.)
3. Mahlow, C., Ulasik, M. A., Tuggener, D. (2022). Extraction of transforming sequences and sentence histories from writing process data: a first step towards linguistic modeling of writing. Read Writ.
4. Kibrik, A. A., Podlesskaya, V. I. (eds.). (2009). Rasskazy o snovideniyah: Korpusnoe issledovanie ustnogo russkogo discursa = Stories about dreams. A Corpus study of Russian oral discourse. Moscow: LRC Publishing House. (In Russ.)
5. Cislaru, G., Olive, T. (2018). Le processus de textualisation. Analyse des unités linguistiques de performance écrite. Louvain-la-Neuve: De Boeck Supérieur. (In French).
6. Baaijen, V. M., Galbraith, D., de Glopper, K. (2012). Keystroke analysis: Reflections on procedures and measures. Written Communication, 29(3), 246–277.
7. Conijn, R. et al. (2022). A product-and process-oriented tagset for revisions in writing. Written Communication. 39(1), 97–128.
8. Miletic, A., et al. (2022). Pro-TEXT: an Annotated Corpus of Keystroke Logs. In Proceedings of LREC (pp. 1732–1739).
9. Goodkind, A. (2021). TypeShift: A User Interface for Visualizing the Typing Production Process. arXiv preprint arXiv:2103.04222
10. Vandermeulen, N., Steendam, E. Van, Rijlaarsdam, G. (2023). Writing Process Feedback Based on Keystroke Logging and Comparison with Exemplars: Effects on the Quality and Process of Synthesis Texts Written Communication, 40(1), 90–144.
11. Meulemans, C. et al. (2022). Cognitive Writing Process Characteristics in Alzheimer's Disease. Frontiers in psychology, 13. 10.3389/fpsyg.2022.872280.
12. Holmlund, T. B. et al. (2022). Towards a temporospatial framework for measurements of disorganization in speech using semantic vectors. Schizophrenia research. 10.1016/j.schres.2022.09.020.
13. Heinen, D. J. P, Johnson, D. R. (2018). Semantic distance: An automated measure of creativity that is novel and appropriate. Psychology of Aesthetics, Creativity and the Arts, 12(2). 10.1037/aca0000125..
14. Beaty, R. E., Johnson, D. R. (2021). Automating creativity assessment with SemDis: An open platform for computing semantic distance. Behavior research methods, 53(2), 757–780.
15. Litvinova, T. (2020). Process-oriented characteristics of an idiolect for authorship attribution of heterogeneous texts: A pilot study. CEUR Workshop Proceedings, 2780, 3–16.

ИНФОРМАЦИЯ ОБ АВТОРЕ

Литвинова Татьяна Александровна

доктор филологических наук
ведущий научный сотрудник научно-исследовательской лаборатории компьютерной семасиологии
Воронежского государственного педагогического университета

INFORMATION ABOUT THE AUTHOR

Litvinova Tatiana Aleksandrovna

Doctor of Philology (Dr. habil.)
Leading Researcher in Computer Semasiology Laboratory, Voronezh State Pedagogical University

Статья поступила в редакцию	24.04.2023	The article was submitted
одобрена после рецензирования	20.05.2023	approved after reviewing
принята к публикации	03.07.2023	accepted for publication