

Бранко Тошович

ТЕКСТЫ, ПОРОЖДАЕМЫЕ СЕТЕВЫМИ ГЕНЕРАТОРАМИ

Данное исследование состоит из трех частей: в первой рассматриваются лингвистические аспекты искусственного порождения текстов, в частности по отношению к естественной генерации текстов, во второй проводится анализ (с привлечением конкретных примеров) основных генераторов, структуры и типов текстов, порождаемых/порожденных сетевыми генераторами, в третьей в центре внимания находятся сетевые генераторы славянских языков. Из-за большого объема статьи здесь публикуются в сжатой форме первая и вторая части, а третья пропущена и будет напечатана позднее в другом месте. Список литературы является максимально селективным (особенно для первой части), ссылки на источники и примеры носят иллюстративный характер, а анализ проводится на материале одного (самого крупного) славянского языка (русского).

0. Генератором является программа, автоматически порождающая на базе одного текста (T_1), или нулевого текста (T_0),¹ другой текст (T_{2gen}), использующая при этом процедуры автоматизации, анализа и синтеза, формулы, условно-случайные поля в различных целях: утилитарных (для улучшения и ускорения общения, обмена информацией, увеличения рейтинга...), игровых, стилистических, развлекательных (авто)презентационных, демонстрационных и т. п.² Натуральная среда для генераторов – это интернет. Они функционируют, как правило, онлайн, реже офлайн (загружаются из Сети и запускаются с компьютера). Порождаемые тексты отличаются гетерогенностью создателей и пользователей – ими могут быть и

¹ Имеется в виду генерация с чистого листа и при помощи иллюстраций без вербального знака.

² Некоторые авторы толкуют генераторы более узко – только с точки зрения поиска в интернете: „Генератор текста – компьютерная программа, способная генерировать последовательности символов, внешне похожие на текст, но при этом, как правило, лишённые смысла. Такие тексты не представляют никакой ценности для пользователей поиска“ (Павлов, Добров 2009, 312).

профессионалы, и любители. Некоторые программы находятся в открытом доступе, другие являются платными.

1. Главный, основополагающий критерий для большинства порождаемых текстов – уникальность (генераторы предназначены в первую очередь для того, чтобы создавать неповторимые тексты).

2. Генерация (G_t) предполагает наличие факультативного генерирующего (входного) текста (T_1) и обязательного сгенерированного (выходного) текста (T_{2gen}). Она охватывает три пространства: пространство локации T_1 (позиторий), пространство генерации T_2 (депозиторий) и пространство дистрибуции T_2 (репозиторий). Исходный T_1 (исходник) находится в позитории (шаблоне, бланке, таблице, базе данных). При помощи различных процедур в депозитории T_1 модифицируется, трансформируется в T_{gen2} (G_{proz}), а потом перекачивается, откладывается в репозиторий (на сервер, в электронную почту, социальную сеть и др.). Порождение состоит из трех звеньев: $T_1/T_0 \rightarrow G_{proz} \rightarrow T_{gen}$. В процессе генерации взаимодействуют два текста: мотивирующий (T_1) и мотивированный (T_{2gen}). Первый является производящим, второй – производным. Входным текстом (T_1) является один текст (в том числе и нулевой – T_0), выходным (T_{2gen}) – один текст или несколько текстов (от двух до миллиона, даже и более). Поэтому приведенную формулу следуют уточнить: $T_1/T_0 \rightarrow G_{proz} \rightarrow T_{2gen[a,b,c\dots x]}$. В G_t наблюдаются следующие закономерности: (1) чем меньше слов в T_{gen1} , тем легче генерировать $T_{2gen[a,b,c\dots x]}$, (2) чем выше порог предсказуемости конститuentов T_1 , тем легче породить $T_{2gen[a,b,c\dots x]}$, (3) чем меньше экспрессии, эмоции, семантических и стилистических сдвигов в T_{gen1} , тем легче получить осмысленные $T_{2gen[a,b,c\dots x]}$. Из T_1 создается множество $T_{2gen[a,b,c\dots x]}$, но наоборот реже бывает (чтобы из множества T_1 был создан один T_{2gen})³. В функции T_1 может выступать прецедентный текст.

3. В исследовании текстов порождаемых/порожденных сетевыми генераторами ($TxIntGn$) важными вопросами являются (1) общелингвистические и лингвоприкладные аспекты $TxIntGn$, (2) природа $TxIntGn$ в отдельных группах языков (славянских, германских, романских и др.), (3) возможности компьютерной лингвистики в $TxIntGn$, (4) роль и место электронных корпусов в $TxIntGn$, (5) ресурсы интернета как коммуникативной среды и платформы для $TxIntGn$, (6) использование языковых механизмов для $TxIntGn$, (7) соотношение языкового стандарта и субстандарта в $TxIntGn$,

³ Сюда относятся компиляционные тексты – коктейль-тексты, состоящие из смеси слов различных текстов, смешанные тексты, мешанина кейвордов, копипаст-тексты (скопированные и вставленные тексты).

(8) соблюдение литературных норм в TxIntGn, (9) языковые инновации в TxIntGn, (10) роль TxIntGn в теории перевода и др.⁴ Здесь будут рассмотрены некоторые из вышеуказанных вопросов. Для этого использован индуктивный метод: сначала была составлена база данных доступных генераторов, затем была предпринята попытка обнаружения их общих свойств и возможностей и определения главных признаков TxIntGn. К исходным пунктам относятся следующие: 1. порождение текстов искусственного языка (ArtifGen) значительно отличается от порождения текстов естественного языка (NatGen), 2. функции TxIntGn характеризуются по отношению к функциям текстов естественного языка (TxNatGn) редукционизмом, 3. глобальная дифференциация естественного языка сужается в TxIntGn на некоторые разновидности.

4. В анализе ArtifGen надо различать два понятия: модель и генератор, моделирование и генерацию/порождение. Моделирование является конструированием другого (реального или воображаемого объекта, изоморфного данному в каких-то существенных признаках; Леонтьев 2003, 41), а генерация/порождение – превращение при помощи моделирования абстрактной системы в конкретную: текстуальную. Модель – это система, которая, отображая или воспроизводя объекты исследования, способна замещать его так, что ее изучение дает новую информацию об этом объекте (Леонтьев 1974, 36). Генератор – это инструмент(арий), механизм, автомат (**A**), который при помощи формулы, алгоритма, схемы, конструкции, модели и любой другой формализованной структуры (**B**) порождает текст (**B**). Во взаимодействии **A** и **B** существует облигаторная связь (**A** не может без **B**) и факультативная связь (**B** может без **A**). Генератор подразумевает три члена (звена): исходный языковой материал (**a**) – средство (**b**) – результат: конечный языковой материал (**b**).

Подобно тому как не следует рассматривать конечный автомат (нервную сеть – совокупность конечного числа нейронов, в которой каждое окончание одного нейрона связано с телом не более чем одного другого

⁴ Приведенные вопросы относятся к блоку „1.6. Интердисциплинарное исследование славянских языков“ Тематики XVI Международного съезда славистов в Белграде 2018 г., в первую очередь к его пунктам: 1.6.7. Компьютерная лингвистика, 1.6.8. Корпусная лингвистика, 1.6.9. Интернет-ресурсы славянских языков, 2.1.4. Славянские языки и культуры в эпоху информационных технологий, и частично к 1.5.3.1. Онлайн словари славянских языков, 1.7.5. Интернет-коммуникация и проблемы современной нормы в славянских языках, 1.7.4. Инновации в нормах литературного языка, 1.3.10. Информационные диалектные фонды, 2.5.2. Процесс перевода: лингвистические и иные особенности передачи культурных кодов.

нейтрона) как модель, соответствующую некоторой теории центральной нервной системы (Джордж 1963, 126), модель (лингвистическую) нельзя отождествлять с генератором (лингвистическим). Конечный автомат (мозг) порождает естественный текст, а сетевой генератор – искусственный текст (подражающий естественный). Разница между ними и в том, что конечный автомат представляет собой первичный автомат, а генератор – вторичный.

Во взаимодействии **A** и **B** в NatGen и ArtifGen различную позицию занимают способы формализации – формулы, алгоритмы, схемы, конструкции, модели, среди которых выделяются: стохастические модели (порождение цепочки элементов, каждый из которых имеет собственную вероятность), грамматики с конечным числом состояний (закономерности распределения единиц высших уровней учитывают закономерности распределения единиц низших уровней), модели непосредственно составляющих (базируются на деривации – последовательной подстановке на место более крупной единицы двух компонентов, из которых она состоит, в результате чего создается „терминальная цепочка“), модели, основанные на трансформационной грамматике (включают три компонента: семантический, фонологический и прагматический), когнитивные модели (существуют различные подходы), психолингвистические модели (выделяются разработки в рамках Московской школы) и др. Собранный нами материал указывает на то, что ArtifGen базируется в настоящее время, как правило, на первой модели – схоластической.⁵ Она пользуется наибольшей популярностью и является самым распространенным, а также самым простым способом порождения.⁶ Ее недостатки состоят в большой нагрузке на процессы, долгом времени генерации и в определенных случаях в порождении текстов без

В качестве основной единицы стохастических процессов выступает не отдельный элемент (фонема, слово...), а определенная последовательность элементов (скажем, цепочка из пяти слов, причем каждое слово имеет вероятность появления, зависящую от появления всех предыдущих). Применимость марковских процессов носит весьма ограниченный характер на собственно грамматическую структуру, так как она не поддается описанию при помощи марковской модели любого типа (Леонтьев 2003, 57–61). Скажем, программа Delirium 1.8 генерирует смешные тексты на основе существующих (используя алгоритм марковских цепочек) и создает довольно длинные осмысленные фразы (Delirium-www). Другой пример – генератор твиттов на Python (Markov Model_Tweets-www). Третий – использование Марковских цепей для обнаружения поискового спама (Павлов, Добров 2009).

⁶ „[...] берется начальное слово и помещается в текст-результат, затем оно ищется в исходном тексте, когда находится, то в текст-результат переносится следующее за ним слово, и уже его ищут в исходном тексте [...]“ (Авт_порожд_текста-www).

смысла („бредотекстов“) или почти без него (Ген_Марк_цепи-www).⁷ Так как потенциал марковских цепей оказывается исчерпанным, предпринимаются попытки по-другому использовать этот метод. Например, Джейми Брю создал симулятор, генерирующий связный текст в зависимости от выбора пользователя (Созд_свою_ист-www), а Тейлор Берг в рамках проекта „Darwin“ предложил пользователям сообща создавать поэзию с помощью процедур, которые имитируют работу генетических алгоритмов (Стохастические стихотворения-www). На принципе случайности базируются SIMP-таблицы.⁸

5. Что касается соотношения ArtifGen и теории языка, надо иметь в виду, что теоретическое языкознание, обобщая и гипотетически осмысляя факты эмпирических разделов лингвистики (диалектологии, истории языка, описательной фонетики, лексикологии и грамматики), а также создавая теоретические концепции языка и текстообразования, осуществляет второй этап лингвистического познания, а на долю раздела лингвистики, занимающегося теорией построения и машинной реализацией моделей языка и текста (инженерной лингвистики), выпадает задача реализовать третий этап лингвистического познания, т. е. проверить на практике состоятельность тех гипотез теоретического языкознания, которые были заложены в воспроизводящие модели инженерной лингвистики (Пиотровский 1979, 12).⁹

TxIntGn (а) показывают, насколько возможно при помощи программных средств автоматически создавать осмысленные тексты в целях ускорения и улучшения коммуникации, экономии времени и усилий, (б) иллюстрируют, в какой степени современные технические и программные средства годятся для сложных лингвистических задач ArtifGen, (в) открывают новые пути развития языка и языкознания. Лингвистические цели TxIntGn

⁷ „У текста, сгенерированного с помощью Марковских цепей, безусловно нет никакой литературной ценности и пользователям он не интересен. Однако, в качестве технического решения может вполне использоваться“ (Марковские цепи-www).

⁸ На таблицах А, В, С, D написаны начала предложений, внутренние части и завершения; из каждой из них выбирается случайный элемент и получается законченное предложение (Авт_порожд_текста-www).

⁹ „Среди всех приложений теоретического языкознания только инженерная лингвистика, оперирующая воспроизводящими языковые и речевые объекты моделями, способная дать однозначную оценку эффективности той или иной лингвистической теории“ (Пиотровский 1979: 12). Инженерно-лингвистическое моделирование служит надежным способом реализации лингвистических теорий и одновременно средством проверки их строгости и жизнеспособности, обнажает скрытые от прямого наблюдения теоретические парадоксы и намечает пути их разрешения (Пиотровский 1979, 13).

зависят от характера генерации. В одних случаях речь идет о более простых интересах – автоматически создавать элементарную языковую единицу (слово, словосочетание, простое предложение), заменять несложную языковую единицу **A** (букву, звук, морфему, слог) такой же единицей **B**, менять их места и т. п. В других случаях в основе лежат более серьезные намерения, и более крупные языковые единицы (сложные предложения) и структуры.

Тексты, относящиеся к ArtifiGen, должны ответить на ряд лингвистических вопросов, например, (1) можно ли при помощи современных программных средств генерировать осмысленные коммуникативные языковые структуры, которые (а) ничем или почти ничем не отличаются от тех, которые создаются человеком без посредничества вспомогательных технических средств и устройств, (2) могут ли TxIntGn обладать функциональностью как (или почти как) их естественные корреспонденты, (3) могут ли они в будущем заменить (или даже вытеснить) классические формы и результаты речевой деятельности, (4) каким образом языковой дуализм (наличие письменной и устной речи) представлен в ArtifiGen, в частности насколько устная речь поддается автоматической генерации. Наблюдаются две закономерности в ArtifiGen: а) число TxIntGn постоянно увеличивается, б) их тематический, содержательный и функциональный круг все больше расширяется, г) в процесс генерации чаще включается то, что легче поддается автоматизации.

6. Генерация текстов с каждым днем совершенствуется, и иногда трудно отличить текст, созданный машиной/программой, от текста, созданного живым человеком. Лингвистическая задача порождения смысловых и коммуникативно пригодных текстов находится в центре внимания компьютерной (инженерной) лингвистики, математической лингвистики и интернет-лингвистики. Этому способствуют сетевая инфраструктура, предоставляющая необходимые механизмы для использования и функционирования TxIntGn.

Для общего языкознания является важным вопрос, как и насколько в TxIntGn реализуются функции языка и в чем своеобразие их проявления. Наши наблюдения указывают на то, что они зависят от типа генераторов и целей TxIntGn. Если исходить лишь из трех элементарных функций: общения, сообщения и воздействия, то оказывается, что в одних случаях на первый план выступает общение, в других – сообщение, в-третьих – воздействие. Но часто в одном и том же тексте переплетается несколько функций. ArtifiGen проявляет самые большие трудности в реализации эстетической функции („сообщении ради самого сообщения“ – Роман Якобсон), которая с трудом поддается любой автоматизации. В основе самого худо-

жественного приема находится дезавтоматизация, нарушение предсказуемости, сдвиг значения, эффект обманутого ожидания, возникновение затрудненной формы, увеличивающей сложность и долготу восприятия. Не простой является и реализация эмоционально-экспрессивной функции. В ряде TxIntGn на первый план выступают функции, которые вообще или почти не упоминаются в теории функций, напр., автопрезентационная и самовыражающая. Для них разработаны сетевые генераторы, при помощи которых проводится внешняя актуализация внутреннего состояния, выражается материальный мир с учетом индивидуального мировоззрения и мировосприятия.¹⁰ Здесь можно добавить и развлекательную функцию, а также психотерапевтическую (поднятие настроения, снятие стресса, эмоционального напряжения, устранение страха различного происхождения, комплексов, особенно неполноценности).¹¹

7. На социолингвистическом уровне ArtifGen вносит свой вклад в новый тип двуязычия – естественно-искусственное.¹² TxIntGn создают конкуренцию естественным текстам, вступают с ними в противоборство и в некоторых областях вытесняют их, особенно если по качеству приближаются к ним.

8. Актуальным является вопрос, насколько глобальная дифференциация языка отражается в TxIntGn и существуют ли определенные ограничения. ArtifGen отличается функционально-стилистическим расслоением, в первую очередь книжных TxIntGn, имеющих официальный характер, и идеолектальных (индивидуальных способов выражения). Все остальные типы членения языка (территориальное, социальное, гендерное, возрастное и др.) находятся пока на периферии ArtifGen. Здесь наблюдается определенная закономерность: чем больше языковые «лекты» удаляются от литературной нормы, тем меньше они поддаются ArtifGen.

9. Важным аспектом TxIntGn является лингвистическое прогнозирование: (а) оценка закономерностей функционирования TxIntGn в будущем, (б) обоснованное суждение предполагаемых изменений, перспективных, потенциальных состояний и альтернативных путей ArtifGen, (в) формирование суждений о последующих состояниях ArtifGen, (г) исследование конкретных перспектив развития ArtifGen отдельных языков. В общем, можно сказать, что ArtifGen не имеет аналога в историческом развитии

¹⁰ Ср. Генератор люб. писем-www.

¹¹ Автопохвалитель-www, Генератор смешн. текста-www.

¹² Об этом понятии см.: Степанов, Панфилов 1983, 228.

языков мира и что ее современное состояние подсказывает несколько возможных сценариев в будущем: 1) число генераторов TxIntGn будет все больше увеличиваться, 2) объекты ArtifGen будут расширяться во всех разновидностях языка, 3) TxIntGn будут все больше заменять и (в некоторых случаях, частично или полностью) вытеснять тексты NatGen, 4) качество ArtifGen будет постоянно приближаться к качеству NatGen, пока в близком или далеком будущем они не сравняются,¹³ 5) чем больше времени будет проходить, тем больше ArtifGen будет в состоянии выполнить (почти) невыполнимую в настоящее время задачу: порождать полноценные метафоры (в широком смысле), создавать разумную экспрессию и совсем осмысленные тексты, в том числе и художественные.

10. Одним из аспектов ArtifGen является ее психолингвистический компонент, а именно восприятие TxIntGen пользователями, наблюдателями и исследователями. Существуют положительные и отрицательные реакции на такие тексты. Положительный отклик вызывают TxIntGn, способствующие общению, ускоряющие коммуникацию и упрощающие обмен информацией. Самый отрицательный эффект порождают бессмысленные, бредовые тексты, небылицы, нонсенсы. Во многом восприятие TxIntGn зависит от успешности ArtifGen (идентификации, сегментации, глубины распознавания, способности декодирования, понимания).

11. Существуют два типа ArtifGen: монолингвальное (в пределах одного языка) и мультилингвальное (в пределах нескольких языков). Первое замыкается в рамки одного языка, в то время как второе представляет собой трансформацию текста T_1 языка L_1 в текст $T_2, T_3 \dots T_x$ языка $L_2, L_3 \dots L_x$ и отличается тем, что всегда предполагает наличие как минимум двух текстов: подлинника и его перевода (в монолингвальном ArtifGen наличие мотивирующего текста на одном и том же языке является факультативным). Мультилингвальное ArtifGen может быть синхронным (мгновенное превращение текста T_1 языка L_1 в текст T_2 языка L_2) и последовательным (сначала переводится текст T_1 языка L_1 на текст T_2 языка L_2 , потом текст T_1 языка L_1 на язык L_3 текста T_3 и т. д.). Мультилингвальное ArtifGen рассматривается в рамках теории перевода, анализа машинного перевода (осуществляемого специальной компьютерной программой)¹⁴ и автомати-

¹³ Уже в настоящее время в некоторых областях машина превосходит человека. Например, робот-шахматист в состоянии выиграть партию у чемпиона мира.

¹⁴ В основном используются два метода: статистический, основанный на сравнении больших объемов языковых пар, и алгоритмический, основанный на правилах.

зированной перевода (осуществляемого человеком при помощи программы). В мультилингвальном ArtifGen важную роль играет „память переводов“ (накопитель переводов, translation memory) – база данных, содержащая набор ранее переведенных сегментов текста (одна запись соответствует одному предложению, реже абзацу).¹⁵ Для некоторых типов перевода разрабатываются специальные лингвистические модели.¹⁶ Мультилингвальные генераторы выступают в форме онлайн переводчиков.¹⁷

12. ArtifGen касается всех языковых уровней. Каждый из них отличается своеобразием и своими возможностями.

На уровне графики разрабатываются сетевые генераторы, которые в состоянии трансформировать текст на одном письме в текст на другом. Здесь существуют различные комбинации: а) текст на буквенном письме **A** (напр., кириллице) ↔ текст на буквенном письме **B** (напр., латинице), б) печатный текст → рукописный текст,¹⁸ в) печатный текст → мультимедийный (анимационный) текст, г) текст на буквенном письме (напр., кириллице) ↔ текст на пиктографии (письме, знаки которого представляют собой схематические рисунки, изображающие предметы действительности и явления природы), д) текст на буквенном письме (напр., кириллице) ↔ текст на идеографии (изображение не предмета, как в пиктограмме, а содержания, значения слова: египетское, китайское письмо), е) текст на буквенном письме (напр., кириллице) – текст на слоговом письме (скажем, индийском), ж) текст на фонографии (слоговом или буквенном письме) ↔ текст на идеографии (египетском, китайском письме).¹⁹ Существует еще одна возможность: автоматически превращать обычную систему письма в специализированную: стенографию (в целях ускорения процесса письменной фиксации живой речи), транскрипцию (письменную фиксацию звучания – фонетическую, фонологическую, научную, практическую), транслитерацию (передачу звучания другого языка буквами своего алфавита). Со-

¹⁵ Если они полностью совпадают, то автоматически вносится перевод. Для памяти переводов разработан международный стандарт хранения данных – ТМХ (Translation Memory eXchange format).

¹⁶ Например, для автоматического перевода научно-технических текстов (рефератов, статей, заголовков патентов) с английского языка на русский создана концепция „Электронно-Технический Автоматический Перевод“ ЭТАП-2 (см.: Лингвистическое обеспечение системы ЭТАП-2 – Добрушин 1989).

¹⁷ Ср.: Яндекс_Переводчик-www, Google Übersetzer-www, Translate.Eu-www, Translate.Ru-www и др.

¹⁸ Сюда относятся генераторы прописей, при помощи которых можно перевести печатный текст в рукописный (Генератор прописей-www).

¹⁹ См.: Искажение текста-www.

ответственно, открываются возможности для автоматического превращения устного текста в его транскрипционную, транслитерационную и стенографическую версии.

Что касается единиц графики (букв, знаков алфавита), то генераторы могут в тексте 1) менять начертание букв (скажем, курсив ↔ полужирный, дизайн, стиль), 2) использовать строчные буквы вместо прописных и наоборот, 3) выделять заглавные буквы и т. п.²⁰ Каждое предложение может начинаться с одной и той же буквы.²¹

ArtifGen на уровне графики не затрагивает содержание, значение, смысл, а лишь меняет, модифицирует форму, в результате чего возникают тексты-инверты, тексты-перевертыши (зеркальные), тексты-метатезы (с заменой букв, слов, строк), анаграммы (с перестановкой элементов), декоративные, дизайнерские, анимированные тексты, „хакерские“, анафорические тексты (все слова начинаются с одной и той же буквы), штрихкодовые тексты, криптографические тексты и др.²² Такие генераторы не требуют сложных моделей формализации.

На фонетико-фонологическом уровне центральной проблемой является вопрос, смогут ли сетевые генераторы порождать полноценные устные тексты.²³ Для этого существуют благоприятные условия, так как в рамках структурного и прикладного языкознания разработаны и интенсивно разрабатываются соответствующие инструментарии. Здесь центральную оппозицию образует пара письменная речь ↔ устная речь. Процесс ArtifGen может идти в оба направления: 1) письменный текст → устный текст, 2) устный текст → письменный текст. В первом случае речь идет о превращении только письменного текста в устный. Один из аспектов генерации письменный текст → устный текст касается возможности трансформации первого во второй путем чтения. Что касается этого направления, генератор должен состоять из двух автоматов: анализатора и синтезатора. В этих целях разработаны голосовые генераторы и генераторы, которые в состоянии прочитать любой текст.²⁴ Во втором возникают две возможности: а) трансформация устного текста в письменный, б) порождение супрасегментарных элементов (ударения, интонации, слога, в первую очередь) и их

²⁰ Напр., Блестящий текст-www.

²¹ Генератор Курочка Ряба-www.

²² Скажем: Текст-инверт-www, Переворачивалка-www.

²³ Фонетический уровень (устный ввод и вывод) наряду с семантическим вызывает основные трудности, испытываемые машиной при обработке естественного языка (Раскин 1971, 261).

²⁴ Одни генераторы предназначены для чтения вслух текстовых файлов (Балаболка-www), другие для чтения по складам (Чтение по складам-www), третьи для голосового набора текста (Голосовой набор-www).

фиксация в письменном тексте. Анализатор осуществляет распознавание (идентификацию), верификацию и анализ речи. Синтезатор базируется на принципах синтеза речи (восстановления формы речевого сигнала по его параметрам) и выполняет различные функции, в том числе озвучивание.²⁵

13. На словообразовательном уровне порождается не целостный текст, а отдельные деривационные единицы. Для ArtifGen особенно интересны разработки в рамках аппликативного порождающего анализа словообразования в 70-ые и 80-ые гг. XX в, направленного на решение трех основных задач: 1) объективации словообразовательной структуры слова, 2) установления соответствий между словообразовательной и морфологической структурой слова, 3) системного представления словообразовательного уровня языка в двух типах языков: генотипическом и фенотипическом (первый моделирует только план содержания, второй – план выражения; Соболева 1972, 165).

14. Для ArtifGen семантика является одним из важнейших уровней языка. Она очень хорошо изучена, и разработан целый ряд моделей значения в различных лингвистических направлениях – структурализме, функциональной грамматике, порождающей грамматике, когнитивной лингвистике. Для ArtifGen особенно интересны: интерпретационная порождающая (генеративная) Н. Хомского, порождающая (генеративная) семантика Дж. Лакоффа, Дж. Мак-Коли, П. М. Постала, модель перехода от смыслов к текстам (языковой синтез) И. А. Мельчука, Ю. Д. Апресяна и А. К. Жолковского (модель Смысл ↔ Текст), модель в форме семантических сетей, модель семантического кодирования и др.

Важнейшим семантическим вопросом для ArtifGen является порождение осмысленных TxIntGn, поэтому смысл (значение, которое слово получает в данном контексте, в данной речевой ситуации) является доминантой в этом процессе. К сожалению, для сетевых генераторов это пока сложнейшая и порой непреодолимая проблема.

15. На лексическом уровне особый интерес вызывает синонимизация. Она является важным приемом ArtifGen. Для этого используются синонимизаторы (синонимайзеры, размножители контента, рандомизаторы, генераторы уникальных текстов) – программы для автоматического порождения рерайт-текстов при помощи синонимов из базы данных.²⁶ Такие генераторы предназначены для веб-мастеров, копирайтеров, владельцев сайтов

²⁵ Сорокин 1992, 12–13.

²⁶ Например, Синонимизатор Intop-www.

и др. в целях наполнения сетевых ресурсов уникальным контентом, прогона сайтов по каталогам статей, регистрации в каталогах сайтов и т. п. Эти инструменты позволяют на базе общего текста (в котором заменяются слова на синонимы, а предложения и абзацы на подобные, похожие по смыслу) генерировать несколько десятков или сотен его вариантов.²⁷ Большинство таких программ используют специальные шаблоны, в которые вносятся исходный текст и потом отдельные слова заменяются при помощи тегов перестановки {}П (Ген.Син Linksfarm-www). В других случаях можно без шаблона (тегов перестановки для пользователя) получить различные варианты одного и того же текста. ArtifGen оперирует понятием случайного (рандомного) слова (неологизма),²⁸ для порождения которого создаются и отдельные генераторы,²⁹ т. е. рандомизаторы.³⁰ Важными являются и ключевые слова (выступают опорными элементами в порождении и восприятии текста), поэтому разрабатываются генераторы ключевых слов,³¹ которые помогают проверить, соответствует ли написанный текст ключевым словам, в целях получения более высокого рейтинга в поиско-

²⁷ Некоторые генераторы работают офлайн (в редакторе Word) благодаря соответствующему макросу (Синонимайка-www) и всплывающему меню. Но такие программы автоматически не синонимизируют выделенный текст.

²⁸ „Метод случайных данных включает в себя сознательное введение чего-то, не связанного с ситуацией. На практике самый простой вид случайных данных это случайное слово. Такое слово можно выбрать действительно случайным образом, используя таблицу случайных чисел и словарь, но есть методы проще, например, ткнуть пальцем в газету и выбрать ближайшее существительное. Случайное слово выступает как связка понятий, вносимых в ситуацию, чтобы открыть новые ходы мысли“ (Случайные данные-www).

²⁹ См.: Генератор случайного слова-www.

³⁰ Рандомизаторы (а) позволяют с помощью синонимов и перестановки создавать уникальный текст (Рандомизатор текста-www), (б) представляют собой рерайтерский инструмент для поточного создания псевдоуникальных текстов (Рандомизатор текста-www2), (в) сочетают черты генератора случайного текста и синонимизатора (Рандомизатор-www).

³¹ Один из них помогает в создании списков ключевых слов при составлении семантического ядра: „Работа этого Генератора проста: Вы создаёте два списка ключевых слов, которые затем комбинируются в один. Например, «купить» и «заказать» в один список, а наименование продуктов в другой – «красную книгу», «зелёную книгу» и т. д. Результат комбинирования будет следующим: *купить красную книгу, заказать красную книгу, купить зелёную книгу, заказать зелёную книгу*“ (Генератор ключ. слов-www). Другие помогают выполнить генерацию потенциальных ключевых слов, взяв за основу указанный текст (будут выбраны наиболее весомые слова текста) – Генератор ключ. слов с текста-www.

вых системах (Seot-www).³² Оптимизатор текста позволяет определить плотность ключевых слов и фраз в тексте, производит автоматический поиск наиболее часто встречающихся фраз (SeoОптимизатор-www).

16. В процессе ArtifGen учитывается тот факт, что морфологические категории закреплены за словами независимо от их синтаксического положения (позиции в предложении) и отвлечены от конкретного высказывания. Так как морфологические формы не могут самостоятельно образовать текст, ArtifGen ограничено на: (а) части речи, (б) структурные элементы слова (морфемы, в первую очередь окончания), (в) морфологические варьирования, (г) категориальные признаки (род, число, падеж, залог, время, вид и др.). Хорошей теоретической основой для ArtifGen являются модели общей морфологии естественных языков и модели для конкретных языков. Здесь является важным процесс нахождения основы слова для заданного исходного слова (стемминг), для чего используются стохастические алгоритмы и строятся вероятностные модели,³³ а также лемматизация (приводящая формы к каноническому виду) и морфологические словари.

17. Предложение является первой языковой единицей, обладающей предикативностью (соотнесенностью сообщения с временным планом действительности), смысловой и интонационной законченностью, поэтому автоматическое порождение этой синтаксической единицы является ключевым для TxIntGn. Неслучайно из-за ее большого значения существуют разнообразные модели формализации и интерпретации, которые так или иначе можно использовать в ArtifGen. Здесь выделяются модели предложения, разработанные в рамках порождающей грамматики, категориальной грамматики, модели смысл ↔ текст, грамматики составляющих, бесконтекстной грамматики, объединенной функциональной грамматики

³² В генерации ключевых слов и фраз участвуют только первые 1000 слов введенного текста. Для ключевых слов используются лексемы с наиболее высокой плотностью в порядке ее уменьшения по отношению ко всему тексту. „Главное правило – ключевики должны соответствовать содержанию той страницы, где они размещены. Генератор ключевых слов из текста использует для анализа предоставленный вами текст, в результате генерации вы получаете до 18 ключевиков, имеющих наибольший вес“ (Генератор ключ. слов с текста-www2).

³³ Существует два основных типа алгоритмов стемминга – алгоритмы поиска (простой стеммер ищет флективную форму в таблице поиска), алгоритмы усеечения окончаний (не используют справочную таблицу с флективными формами и отношениями корня и формы, а список правил, учитывая форму слова для нахождения его основы) – Стемминг-www.

(ОФГ), прямой контроль грамматики при понимании, систематической грамматики, грамматики расширенных сетей переходов (РСП) и др., а также типов синтаксического анализа сверху вниз и снизу вверх.

18. На стилистическом уровне ArtifGen затрагивает вопросы: а) насколько сетевые генераторы в состоянии порождать экспрессивность, б) смогут ли они справляться с образностью, в) можно ли создавать тексты на базе осознанного нарушения языковых норм, г) как и в какой степени генерируются тексты различных стилей. Что касается последнего вопроса, уже существуют генераторы, порождающие тексты всех функциональных стилей. Легче всего это удастся по отношению к текстам официального характера, стилистически сухим, неэкспрессивным (в первую очередь типовым документам). Сложнее всего с литературно-художественным, но и здесь наблюдаются прорывы.

19. Своеобразие славянских языков – флективный характер,³⁴ наличие большого количества грамматических форм, развитость падежной и видовой систем создает немалые проблемы для ArtifGen.³⁵ Эти языки относятся к синтетическим, в которых грамматическими маркерами являются окончания, приставки, суффиксы, внутренняя флексия, изменение ударения, супплетивные формы, требующие серьезной обработки при ArtifGen и существенным образом усложняющие этот процесс. Для аналитических языков (романских, германских и др.) такие трудности являются не такими большими, так как у них выражение грамматического значения выносится за пределы слова и передается с помощью союзов, предлогов, членов и других служебных слов, а также порядка слов.³⁶

³⁴ „Особенностью славянских языков, как и вообще флективных языков, является то, что отдельные грамматические категории в них не получают непосредственное выражение, а поэтому не могут получить простого определения. Поэтому для них приходится строить достаточно сложные модели“ (Ревзин 1967, 137).

³⁵ Особняком здесь стоит болгарский язык, почти утративший флективность в категории имен существительных.

³⁶ „Для языков с простой морфологией наподобие английского размеры таблиц небольшие, но для сильно флективных языков (например, турецкий) таблица может иметь сотни возможных флективных форм для каждого корня“ (Стемминг-www). Но в переработке аналитических языков, каким является английский, возникает другая трудность: „Английский язык, из-за слабой маркировки семантических и синтаксических отношений между словами во фразе, с точки зрения автоматической обработки по крайней мере на порядок сложнее французского и в особенности русского. Действительно, части речи в нем почти не имеют характерных парадигм, и одно орфографическое слово сплошь и рядом

20. Важной основой для ArtifGen является компьютерная лингвистика (занимающаяся автоматическим анализом естественного языка и моделированием языка с помощью компьютера)³⁷ и ее разделы: компьютерная фонетика (анализ и синтез речи), компьютерная морфология (морфологический анализ и синтез, построение морфологических модулей), компьютерный синтаксис (парсинг)³⁸, компьютерная семантика (моделирование смысла текста), компьютерная лексикография (создание электронных словарей).

21. Корпусная лингвистика предоставляет очень полезный материал для ArtifGen – семантическое, грамматическое (морфологическое, синтаксическое) аннотирование (разметку), лемматизацию, размеченные корпуса и др.

22. ArtifGen порождает различные типы текстов. Прежде всего выделяются входные и выходные тексты.

23. Входные (генерирующие) тексты T_1 могут быть интегральными ($T_{1Integr[a,b,c...x]}$), сегментарными ($T_{1Segm[a,b,c...x]}$) и элементарными ($T_{1Elem[a,b,c...x]}$). К интегральным ($T_{1Integr[a,b,c...x]}$) относятся полноценные рассказы, стихотворения, документы, письма и т. п. Сегментарные тексты ($T_{1Segm[a,b,c...x]}$) представляют собой части $T_{1Integr[a,b,c...x]}$ типа рифм. Элементарными текстами ($T_{1Elem[a,b,c...x]}$) являются ссылки, в которые вставляются сгенерированные анкоры (между $\langle a \rangle$ и $\langle /a \rangle$).

24. Интегральные тексты $T_{1Integr[a,b,c...x]}$ – это готовые тексты. Сюда относятся (а) фейк-источники, (б) кодовые, прецедентные, инхоативные, поисковые, медийные, контентно-объектные, формально-объектные, рандомизируемые (синомизируемые), количественно (не)лимитированные, осмысленные и бессмысленные тексты. Их можно разделить на нерасчлененные

бывает существительным, прилагательным и глаголом, а иногда даже наречием и предлогом [...] Окончания разных частей речи омонимичны [...] Прилагательные и причастия не имеют никаких окончаний, и поэтому их согласование с существительными никак не выражено [...] Подчинительные союзы и союзные слова могут опускаться. [...] Тем самым оказываются невыраженными подчинительные отношения. [...] В характерных для английского языка цепочках существительных синтаксические связи тоже никак не маркируются, и поэтому такие цепочки сами по себе не дают никакого ключа к их правильной семантической интерпретации“ (Добрушин 1989, 14).

³⁷ Она является правопреемницей инженерной лингвистики (Коваль 2005, 44).

³⁸ Речь идет о разборе в форме синтаксическое дерева.

и расчлененные. К группе нерасчлененных принадлежат T_1 , выступающие в классическом виде (в неразрывной последовательности предложений и абзацев): фейк-источники, прецедентные, деструктивно-объективные, формально-объектные, синонимизируемые тексты. Расчлененными являются аналитические T_1 . К таким относятся кодовые, инхоативные, сборочные (на базе результатов поиска), пошаговые, медийные тексты. И те, и другие (нерасчлененные и расчлененные) могут быть количественно (не)лимитированными, а также осмысленными и бессмысленными.

Фейк-источники являются подлинниками T_1 , на основе которых генерируются $T_{2gen[a,b,c...x]}$ с фейковым, фальшивым контентом³⁹ (в том числе и псевдотвитты).⁴⁰ Кодовые T_1 выступают в форме таблицы с штампованными фразами, расставленными по столбцам и в определенном порядке (тексты разбиты на предложения, каждое из которых занимает отдельную строку или отдельный столбец).

Прецедентные T_1 представляют собой общеизвестные тексты в конкретной речевой культуре, на которые часто ссылаются. Инхоативные T_1 используются для того, чтобы создать только начало для $T_{2gen[a,b,c...x]}$; они предоставляют возможность пользователю продолжать и заканчивать сгенерированную инициальную часть.⁴¹ Сборочные (поисковые) T_1 выступают в форме результатов поиска, на базе которых генерируются $T_{2gen[a,b,c...x]}$. Пошаговые T_1 подразумевают последовательный выбор элементов и сегментов $T_{2gen[a,b,c...x]}$.⁴² Медийные T_1 состоят только из вербального или только невербального кодов, т. е. являются мономедийными, а при помощи генератора превращаются в мультимедийные $T_{2gen[a,b,c...x]}$. Такие тексты порождают генераторы (де)мотиваторов. Их позиторий состоит из трех частей: изображения (фотографии, картинка и т. п.), верхней надписи (заголовка), нижней надписи (дополнительного текста). Деструктивно-

³⁹ См. Fake News-www1.

⁴⁰ Такие $T_{2gen[a,b,c...x]}$ порождают генераторы типа „Твитты за Кадирова“ по схеме: T^1 = подлинник (имплицитный) – T^2 = симулякр, напр. *Желаю Магомеду культуры. ♦ По милости Всевышнего Аллаха мы встретились за их лобные коварные научные трубы* (Neugomzan-www).

⁴¹ Скажем, начальную часть стихов порождает генератор, а конечную – живой поэт: *Не жалею, не зову, не плачу, Все пройдет, как с белых яблонь дым. → Понадеешься вдруг на удачу. Вот мы с тобой и шумим* (Стихоробот-www).

⁴² Например, порождение мотиватора проводится по следующим фазам: „1. Задаем ориентацию: книжная или альбомная. 2. Жмите «открыть» и загружайте изображение. 3. Выберите шрифт и его цвет. 4. Текст – будет располагаться под заглавием, более мелкого шрифта. Это любой ваш комментарий к мотиватору. 5. Граница – тонкая рамка вокруг картинка мотиватора. 6. Фон мотиватора. Выбирайте позитивные цвета [...] 7. Сохраните готовый мотиватор [...]“ (Мотиватор-www).

объектные T_1 являются предметом сознательного искажения в целях создания эффекта в $T_{2gen[a,b,c...x]}$.⁴³ Формально-объектные T_1 представляют собой материал для формальной (не содержательной) модификации и превращения в $T_{2gen[a,b,c...x]}$; речь идет о создании нового дизайна, выбора формата, символов и т. п.⁴⁴ Рандомизируемые/синонимизируемые T_1 (размножаемые, тиражируемые, дистрибутивные, ротаторские, трафаретные тексты) используются в целях получения текстов-приманок для поисковиков. Их порождение идет по двум направлениям: а) один и тот же текст тиражируется в многочисленные синонимизированные копии, б) один и тот же текст распространяется по многочисленным адресам в Сети.⁴⁵ Сюда относятся рерайтинг-тексты (тексты, создаваемые заимствованием контента из различных источников, рерайтингом, тексты с плагирированным, воровским контентом). В целях синонимизации используются рандомизаторы текста (напр., Рандомизатор текста-www). Они могут очень быстро из одного текста сделать несколько десятков или сотен его вариантов.⁴⁶

Количественно (не)лимитированные T_1 охватывают (а) исходные тексты, ограниченные в наборе знаков (напр., до 6 000 символов)⁴⁷ и (б) без таких рамок. T_1 могут быть осмысленными и бессмысленными⁴⁸.

25. Сегментарные тексты $T_{1Segm[a,b,c...x]}$ бывают параметрическими, тематическими и жанровыми. Параметрические T_1 выступают в форме элементов

⁴³ Сюда относится сетевая поэзия, возникающая в результате модификации и (иногда) полного искажения T_1 (авторского текста). Типичным примером являются тексты, создаваемые генераторами стихов в стиле А. С. Пушкина. Ср.: T_1 : *Я помню чудное мгновенье: | Передо мной явилась ты, | Как мимолетное виденье, | Как гений чистой красоты.* T_{2gen} : *Не помню странное везенье | В моих мечтах стихов листы | Как мимолетное виденье | Как символ липкой суеты* (Генератор стихов2-www).

⁴⁴ Таким является, например, генератор, преобразующий двоичный код в текст и наоборот (Бинарный код в текст-www).

⁴⁵ Генератор отправки php-www.

⁴⁶ Вот один из них: *1-2 комнатные квартиры в центре. Хороший ремонт, пылесос, 2 телевизора, кабельное ТВ, посуда, микроволновка, пластиковые окна, Интернет, уют, стиральная машина, новая мебель. Все документы. В этой квартире очень тихо, уютно, комфортно. Рядом с домом 3 супермаркета, 2 аптеки, 2 стоянки для автомобилей.*

⁴⁷ Генератор текстов-www.

⁴⁸ Бессмысленные тексты порождают различные генераторы. Вот один пример: *Чей был казнь тон Чтя Враги грянь нашей ком Вот. Соединюся прошедший познавать судилищах склонился богатству. Наполнить праведник раздаётся содержишь действует. Вы им он со Не ль За яд* (Randomtextgenerator-www).

бланков, графов, шаблонов с критериями для порождения $T_{2gen[a,b,c...x]}$. По ним можно и заказывать тексты.⁴⁹ Тематические T_1 имеют форму списка тем, предлагаемых для порождения T_2 .⁵⁰ Жанровые T_1 представляют собой набор жанров для генерации $T_{2gen[a,b,c...x]}$.⁵¹ Сюда относится и стилевая генерация – использование алгоритмов для написания текстов в стиле определенного автора.⁵²

26. Генерируемые тексты $T_{2gen[a,b,c...x]}$ отличаются различной степенью утилитарности (одни из них являются очень полезными, другие же не имеют никакой практической ценности). Среди них выделяются те, которые обладают смыслом, и те, которые являются бессмысленными, нонсенсными (нонсенс-текстами), хотя оба типа характеризует грамматическая правильность. Здесь возникают следующие комбинации: нормативность + осмысленность, нормативность + неосмысленность. В исследуемом нами материале отсутствует связь: ненормативность + неосмысленность. В генераторских текстах наблюдается в одних случаях симметрия (совпадение) нормы и смысла, а в других – их асимметрия. $T_{2gen[a,b,c...x]}$ бывают (литературно) внятыми,⁵³ осмысленными, почти осмысленными, бессмысленными (не имеющими никакого смысла), небылицами.⁵⁴ Некоторые генераторы являются формообразующими – при их использовании

⁴⁹ Большинство таких генераторов используются для составления поздравлений в стихах (Генератор поздравлений-www).

⁵⁰ Существуют генераторы тем рефератов, курсовых и дипломных работ, например, по педагогике: *Жан-Жак Руссо и его отношение к идее свободного воспитания* ♦ *микрокалькулятор как предмет пристального изучения педагогики* (Генератор тем-www). Некоторые генераторы создают не элементарные, сегментарные части текстов (заголовки), а целые тексты на определенную тему типа „Сознание как восприятие“: *Когнитивная составляющая параллельна. Предсознательное последовательно вызывает экспериментальный конформизм [...]* (Яндекс Рефераты-www).

⁵¹ Речь идет о порождении стихов в определенном жанре, скажем, хокку с их жесткой структурой: первая строка – пять слогов, вторая – семь, третья – пять. Ср.: *Устал. Колени дрожат, | Братья друг другу мы все. | Завтра отправлюсь в Китай.* ♦ *Детский доносится смех, | Я поднялся на холм. | Лишь меч не знает греха* (Генератор хокку-www).

⁵² Напр., Алексей Тихонов и Иван Ямщиков выпустили альбом „Нейронная оборона“, который состоит из песен и стихотворений, написанных роботом, а по образцу стихов Егора Летова, основателя группы „Гражданская оборона“ типа: *В ожидании чудес, невозможных чудес | Я смотрю в темноту, но я не верю в прогресс | Я хочу быть убийцей, я хочу быть живым | Мне осталось всего лишь дожить до седин [...]* (Генератор песен ГО-www).

⁵³ Генератор внятных текстов-www.

⁵⁴ Нонсенский генератор-www.

меняется форма T_{gen1} , а содержание не модифицируется.⁵⁵ Другие же являются контентообразующими (они предназначены для порождения $T_{2gen[a,b,c...x]}$ с измененным контентом).⁵⁶ Существуют моноэкземплярные (однотиражные) и мультиэкземплярные (многотиражные, мультипликативные) $T_{2gen[a,b,c...x]}$. Моноэкземплярные (однотиражные) тексты генерируются по принципу $1 \rightarrow 1$: из одного экземпляра T_1 возникает один экземпляр T_{2gen} . Мультиэкземплярные (многотиражные) порождаются по принципу $1 \rightarrow 1, 2, 3... x$: на основе одного экземпляра T_1 создается несколько $T_{2gen[a,b,c...x]}$; при этом T_1 оптимизируются, а потом размножаются.

Что касается качества, $T_{2genElem[a,b,c...x]}$ могут быть шедеврами, а также текстами среднего или низкого качества. Количественная шкала возможностей T_{2gen} очень широка: от одного слова до многостраничного текста.

27. Одними из $T_{2genElem[a,b,c...x]}$ являются провизорные, демонстрационные, неутвержденные тексты, предназначенные для дизайнеров, верстальщиков (разбивающих текст на отдельные страницы, готовящих оригинал-макет издания), копирайтеров (создающих рекламные и презентационные тексты), рерайтеров (обрабатывающих исходные тексты в целях их дальнейшего использования), спамеров (готовящих поисковый спам) – в общем для специалистов более узкого профиля – авторов сетевого материала. Это „до-тексты“/„пре-тексты“/„пред-тексты“, тексты-симулякры (тексты, похожие на тексты; тексты и не тексты; тексты, подобные текстам), тексты-суррогаты (заменяющие другие, настоящие тексты, которые потом приходят на их место). По отношению к реальным, они являются случайными, фиктивными, вымышленными, ложными/фальшивыми/псевдотекстами, неестественными, обманывающими/обман-текстами. Так как они используются для создания макетов, их называют и макетными. Многие из них являются неосмысленными и поэтому нечитабельными/ нечитаемыми/ псевдочитабельными („до-тексты“ вообще и не читаются). Из-за своей абсурдности их считают „бредотекстами“. Такие тексты имеют и образные имена, в том числе с лексемой *рыба* / корнем **рыб-**: *рыбный текст, текст рыбы, текст-,рыба“, рыба-текст, рыбате́кст, рыботекст*.⁵⁷ Реже встре-

⁵⁵ Скажем, на базе T_1 : *сегодня* создается T_2 : [сезогнт](#) (Блестящий текст-www).

⁵⁶ Контентообразующим генератором является, например, „Автопоэт“, порождающий стихи типа: *что такое питание | черепах описание | ренессанс страхования | биатлон расписание* (Автопоэт-www).

⁵⁷ Это тексты типа: *Далеко-далеко за словесными горами в стране гласных и согласных живут рыбные тексты. Вдали от всех живут они в буквенных домах на берегу Семантика большого языкового океана. Маленький ручеек Даль журчит по всей стране и обеспечивает ее всеми необходимыми правилами. [...]* (Blindtextgenerator-www).

чаются метафорические словосочетания: *желтый трамвай, желтый метро, текст манекен, манекен текст, белый текст*. Для номинации используются также иноязычные названия типа *Loren Ipsum*,⁵⁸ *Litlipsum*,⁵⁹ *dummy text, dumm Text*.

„До-тексты“ симулируют, имитируют реальные тексты, не порождаемые генераторами. В G_t существует генерирующий текст (лежащий в основе порождения, T_1), генерируемый текст (мнимый, эксплицитный, демонстрационный, T_{2gen}) и негенерируемый (ориентационный, последующий, модально-имплицитный, но настоящий, целевой – $T_{3orient}$). На первом этапе порождается „до-текст“ (T_{2gen}), затем при его помощи готовится форма последующего, конечного текста – „пост-текста“. В начале процесса неизвестно, каким по содержанию будет целевой, настоящий „пост-текст“ ($T_{3orient}$). Когда обработка и проверка „до-текста“ (T_{2gen}) заканчиваются, он устраняется и заменяется „пост-текстом“ ($T_{3orient}$). Но это происходит за рамками генерации. Для „до-текстов“ смысл не нужен – они выполняют лишь функцию временного наполнителя/заполнителя (являются альтернативными текстами). Поэтому это тексты на лету, по умолчанию. Их преимущество в том, что (1) генерируются они очень быстро, без больших и сложных процедур, (2) не отвлекают внимание разработчиков смыслом. Такие фиктивные тексты предназначены для того, чтобы подготовить реальный текст и продемонстрировать, как он выглядит. „До-тексты“ относятся к нереферативным текстам, так как отсутствует их соотносительность с (реальными) денотатами.

28. Меню-тексты T_2 предоставляют возможность выбирать последовательно (не параллельно) из невидимого меню один текст (T_{2a}) или несколько текстов ($T_{2a,2b,2c...2x}$), причем список является скрытым. Он может быть активным (выпадающим) и пассивным (фиксированным). Для того, чтобы продолжить выбор, если сгенерированный текст не устраивает, надо нажать кнопку типа *Еще* (и т. п.).⁶⁰

⁵⁸ Латинские „до-тексты“ Lorem Ipsum могут быть готовыми, модифицированными, новыми и псевдотекстами. Примером Lorem Ipsum является такой фрагмент: Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed est augue, vestibulum vel ipsum non, luctus mattis risus. Fusce eget enim quis ex suscipit aliquet eu id nunc. Phasellus vitae odio diam (Lorem Ipsum-www).

⁵⁹ Речь идет о маленьких бессмысленных латинских текстах (Litlipsum-www).

⁶⁰ Меню-тексты порождает, например, Генератор объясн. записок-www.

29. Выходные $T_{2gen[a,b,c...x]}$ могут быть, как и T_1 , интегральными ($T_{2genIntegr[a,b,c...x]}$), сегментарными ($T_{2genSegm[a,b,c...x]}$) и элементарными ($T_{2genElem[a,b,c...x]}$).

30. К интегральным $T_{2genIntegr[a,b,c...x]}$ относятся сгенерированные полноценные рассказы, стихотворения, документы, письма и т. п. Они образуют пять групп текстов: мультимедийные (объединяют как минимум два типа медиа) – комиксы, (де)мотиваторы, мемы, мультимедиаальные истории, фоновые (со скрытым текстом) – QR-тексты, графические (тексты только с измененной формой) – тексты-инверты, тексты-перевертыши (зеркальные), тексты-метатезы (с заменой букв, слов, строк), анаграммы (с перестановкой элементов), декоративные, дизайнерские, анимированные, „хакерские“, анафорические (все слова начинаются с одной и той же буквы), , штрихкодовые, криптографические тексты, перлокутивные – игровые, смешные, сатирические, забавные тексты, функционально-стилистические – прозаические: научно-фантастические, сказочные, поэтические (спам-поэзия, песни), научные (рефераты, рукописные конспекты, псевдостатьи – псевдонаучные, псевдоосмысленные, псевдофилолофские тексты, выполняющие иногда провокационную функцию⁶¹), деловые (документы, биографии,⁶² долговые расписки, объяснительные записки,⁶³ отчеты, отзывы, рекламы), политические, (де)мотиваторские, интерперсональные – тексты-поздравления, тексты-тосты, автопохвалительные,⁶⁴ автопрезентационные, комплиментные,⁶⁵ эпистолярные – любовные письма и послания, письма

⁶¹ Таким генератором является SCIGen, который порождает случайный текст, напоминающий научную статью (SCIGen-www).

⁶² *Барицев Илиодор Матвеевич. Родился в 1939 году в городе Тура. Учился в школе №16. Закончил ЕГПТУ по специальности экономист. Работал депутатом ГД РФ. Сейчас на пенсии воспитывает внуков (Генератор биографий-www).*

⁶³ *Сегодня с утра я снова опоздал на работу. Мы взяли бутылку незнакомого вина, а оно оказалось долгоиграющим, отчего я и опоздал на работу. Прошу не принимать в отношении меня административных и других санкций. Спасибо за понимание (Генератор объясн. записок-www).*

⁶⁴ *Вы способны и талантливы! ♦ Вам завидуют! ♦ Вы добрый! ♦ Вы красивый! ♦ У вас все хорошо! По утрам Вам улыбается солнце! (Автопохвалитель-www).*

⁶⁵ **[28]** *Олененок мой Оля! я давно хотел тебе сказать, что если ты грустишь, то на меня находит жуткая хандра... кроме того, у тебя такие нежные глаза (Генератор комплиментов-www).*

от Деда Мороза, светские беседы, тексты для членов жюри КВН,⁶⁶ разговорные – озвученные тексты говорилками, голосовыми генераторами⁶⁷.

31. Сегментарные тексты представляют собой сгенерированные части $T_{2genInteg[a,b,c...x]}$ (напр., рифмы, сниппеты⁶⁸).

32. Элементарными текстами $T_{2genElem[a,b,c...x]}$ являются сгенерированные заголовки имен, ключевые слова, анкоры, афоризмы, случайные слова, названия блюд, названия населенных пунктов, ключевые фразы для текстов ссылок, логотипы, ники, смайлики, пароли, слоганы, значки и т. п.⁶⁹

33. В проведенном анализе были выявлены и (частично) рассмотрены некоторые типы генерации текстов. Они имеют форму моделей, состоящих из пяти элементов: название (образное) модели, тип генерации, формула, номер примера, схема.⁷⁰

Модель	Тип генерации	Формула	Пример
{Напр.: „одуванчик“}	{Напр.: „одуван- чая“}	{Напр.: жанро- вая}	{Напр.: [3]}
		$2gen[a,b,c...x]$	
		Схема	

Таковыми являются: **1.** „чистый лист“ (генерация без исходника, исходного текста), **2.** „зеркало“ (одному исходному T_1 соответствует один сгенерированный T_2), **3.** „одуванчик“ (одному исходному T_1 соответствует несколько сгенерированных T_2), **4.** „коктейль“ (на базе определенных параметров T_1 создается T_2), **5.** „эстафета“ (T_2 получается проведением последовательных процедур), **6.** „наклейка“ (на мономедиа T_1 наклеиваются, накладываются

⁶⁶ *Дорогие друзья! Я хочу отдельно сказать о капитанах команд. Я видел, что вы отдали игре все силы, энергию, время и деньги. Поэтому все вы скоро из обычных капитанов превратитесь в ефрейторов, а, может, и в сержантов. [...]* (Генератор текстов для КВН-www).

⁶⁷ Говорилка-www, Балаболка-www.

⁶⁸ Сниппет (*snippet* ‘отрывок, фрагмент’) – фрагмент исходного текста или кода программы, применяемый в поисковых системах, текстовых редакторах и средах разработки (Сниппет-www).

⁶⁹ Например, один из генераторов порождает такие имена: *Вскооир* ♦ *Лин* ♦ *Сэмлуил* ♦ *луиф* ♦ *Дреб* [...] (Генератор имен и названий-www). Другой генерирует названия публикаций: *Судьба господина* ♦ *Преодолевая Бога* ♦ *Слово героя* [...] (Генератор названий книг-www).

⁷⁰ Некоторые из них относятся не только к одному типу, а к нескольким (двум, трем...).

мультимедиа), **7.** „фундамент“ (T_1 является фундаментом для надстройки в форме T_2), **8.** „лото“ (из различных элементов/частей T_1 монтируется T_2), **9.** „кривое зеркало“ (в левом пространстве зеркала находится подлинник T_1 , а в правом – его искажение в форме T_2), **10.** „рамка“ (T_1 является жанром – объектом подражания и превращения T_2), **11.** „орел – решка“ (форма T_1 модифицируется и трансформируется в T_2), **12.** „меню“ (из скрытого меню всплывает по желанию пользователя один, два, три... T_2), **13.** „подопытный кролик“ (T_2 является экспериментальным, пробным, проверочным для T_3 , не включенного в G_i), **14.** „шапка-невидимка“ (T_2 является скрытым, невидимым T_1), **15.** „энтер“ (T_2 порождается кликом на клавишу „энтер“ после T_1 в форме формулы), **16.** „рукав“ (T_1 – элемент, часть для порождения гипотетических, модальных T_2), **17.** „ключ“ (T_2 получаются кликом на „Еще“ или на подобную команду), **18.** „реферат“ (T_2 выступает в форме сгенерированного названия темы или интегрального текста на определенную тему T_1), **19.** „гребешок“ (каждое предложение T_2 начинается с одной и той же буквы), **20.** „укол“ (T_1 получает „инъекцию“ – новое содержание и превращается в T_2), **21.** „говорящая тишина“ (сгенерированный текст T_2 является бессмысленным, предназначенным только для демонстрации T_3 за рамками генерации), **22.** „тень“ (T_1 представляет собой скрытый текст, т. е. фигуру, дающую тень – T_2).

34. Проведенный анализ указывает на то, что генераторы порождают осмысленные, полуосмысленные, смыслоподобные и бессмысленные тексты. TxIntGn отличаются разнообразной тематикой, различным назначением, неодинаковой практической ценностью и качеством, а также неоднородными целями: утилитарными, игровыми, стилистическими, развлекательными, (авто)презентационными, демонстрационными. Естественное порождение текстов (NatGen) отличается от искусственного (ArtifGen) в частности в том, что (а) NatGen происходит по природным алгоритмам, (б) что ArtifGen требует разработки правил и команд автоматизации и генерации, (в) для ArtifGen используются модели, которые для NatGen являются неполноценными. TxIntGn часто напоминают естественные тексты, бывают похожими на них, поэтому из трех типов отношений естественных и искусственных текстов: тождества ($A = B$), сходства ($A \approx B$) и различия ($A \neq B$) здесь преобладает сходство. Это касается и формы (например, в сетевой поэзии выделяются две категории: настоящие и стихоподобные тексты). TxIntGn все больше проникают в повседневную жизнь и человеческую деятельность. Хорошей основой для ArtifGen являются разнообразные методы и модели формализации NatGen. Они больше ориентированы на анализ, чем на синтез, что является одной из причин, почему они реже используются для ArtifGen. Для создания искусственного интеллекта

в области языка требуется время (такие разработки не уходят своим корнем в далекое прошлое) и больших усилий (творческих и технических). ArtifGen за очень короткий период компьютеризации и интернетизации (полстолетия) не смогло достичь уровня высокого качества и полностью решить сложные вопросы, связанные со смыслом, метафоризацией и экспрессивностью. Сетевые генераторы являются одним из средств проверки на состоятельность лингвистических гипотез и формализованных инструментариев. Это своеобразная лакмусовая бумага для определения годности лингвистических моделей к их практическому использованию. ArtifGen открывает новые пути в развитии языка и его исследованиях.

35. Исходя из результатов приведенного анализа можно сделать ряд прогнозов и гипотез. В ближайшем будущем (маловероятно) или в далеком будущем (более вероятно) качество и функциональность ArtifGen приблизится к NatGen настолько, что очень трудно будет отличать естественные тексты от искусственно порождаемых/порожденных. TxIntGn будут постоянно совершенствоваться и повышать качество и эффективность, что приведет к тому, что во многих сферах жизни и деятельности они будут полностью или частично заменять/вытеснять тексты, создаваемые классическим путем. Экспансия TxIntGn будет иметь необратимый характер. Предметом ArtifGen будет и то, что в настоящее время вообще не поддается генерации. Технологический прогресс в далеком/близком будущем объединит в одно целое два автомата – естественный (человеческий мозг) и искусственный генераторы, так что границы между NatGen и ArtifGen будут во многом нейтрализованы. Этому будет способствовать гибридизация человека, организм которого будет напичкан чипами, электронными сетями, базами данных, комплексами знаний и самыми различными генераторами, в том числе и теми, которые будут порождать осмысленные тексты (трудно ответить на вопрос о том, перестанет ли тогда человек быть человеком, а автомат только машиной). Генераторы будут автономно клонировать самих себя, улучшая и ускоряя общение. Гибридизированный (чипизированный и автоматизированный) человек будет способен создавать тексты, сильно отличающиеся от существующих. Генерация текстов таким симбиозным механизмом (человеком-автоматом) приведет к превращению естественно-искусственного двуязычия в одноязычие, контуры и размеры которого трудно представить в настоящее время. Генераторы текстов настолько будут развиты, что вытеснят многие профессии и виды деятельности (уже в настоящее время ставится вопрос, убивают ли генераторы „автопоэты“ живых поэтов). Художественное творчество будет существенно изменено и расширено тематически, методологически, структурно и эстетически.

Источники

- Автопохвалитель-www: *Автопохвалитель* <http://nepom.ru/usefulness/auto-praise.php> (4.3.2017).
- Автопоэт-www: *Автопоэт* <https://yandex.ru/autopoet/monorim/20/> (1.6.2017).
- Балаболка-www: *Балаболка* <http://www.cross-plus-a.ru/balabolka.html> (8.1.2017).
- Балаболка-www: *Балаболка* <http://www.cross-plus-a.ru/balabolka.html> (1.4.2017).
- Бинарный код в текст-www: *Бинарный код в текст* <http://php-zametki.ru/binary-text.html> (8.1.2017).
- Блестящий текст-www: *Блестящий текст* <http://gifr.ru/glitter/> (9.2.2017).
- Ген_Марк_цепи-www: *Генератор текста на основе Марковских цепей* <http://logloss.ru/markov-text-generator/> (15.4.2017).
- Генератор биографий-www: *Генератор биографий.* <http://ainte.ru/talkbox/person/> (10.2.2015).
- Генератор внятных текстов-www: *Генератор текстов.* <http://kvisaz.livejournal.com/72251.html> (1.6.2017).
- Генератор имен и названий-www: *Генератор имен.* <http://www.klikin.ru/namegen.html> (8.1.2017).
- Генератор ключ. слов с текста-www: *Генератор ключевых слов с текста.* <http://info.seocafe.info/tools/keygeneratortext/> (18.4.2017).
- Генератор ключ. слов с текста-www2. *Генератор ключевых слов с текста.* [2]. <http://webmasta.org/tools/keygeneratortext/> (4.3.2017).
- Генератор ключ. слов-www: *Генератор Ключевых Слов.* <http://www.seotxt.com/service/comb/> (14.2.2017).
- Генератор комплиментов-www: *Генератор комплиментов.* <http://ainte.ru/talkbox/compliment/> (10.2.2015).
- Генератор Курочка Ряба-www: *Генератор текстов на основе патернов, Курочка Ряба и Звездные войны.* <https://habrahabr.ru/post/163727/> (30.5.2017).
- Генератор люб. писем-www: *Генератор любовных писем.* <http://venevchat.narod.ru/love/generator.html> (4.3.2017).
- Генератор названий книг-www: *Генератор названий книг.* <http://www.klikin.ru/genname.html/> (8.1.2017).
- Генератор объясн. записок-www: *Генератор объяснительных записок!* <http://nepom.ru/usefulness/explanation.php> (10.2.2015).
- Генератор отправки php-www: *Генератор формы отправки писем на php.* <http://www.php-mail.ru/?id=6> (1.2.2015).
- Генератор песен ГО-www: *Найден секретный генератор песен гр. ГРажданская Оборона.* <https://soundcoderr.wordpress.com/2016/09/13/13-744/> (1.6.2017).
- Генератор поздравлений-www: *Составить поздравление в стихах.* <http://yubiley.org/sostavit-pozdravlenie-v-stihah.html> (3.2.2015).
- Генератор прописей-www: *[Генератор прописей].* <https://coo.by/writer> (12.1.2017).

- Генератор случайного слова-www: *Генератор случайного слова*. <http://sluchajnoe.ru/slovo.php> (7.6.2017).
- Генератор смешн. текста-www: *Генератор смешного текста*. <https://mvos.ru/post/159> (30.5.2017).
- Генератор стихов2-www: *Генератор стихов* [2]. <http://www.webzona.ru/n/doc/generator-stihov.htm> (3.2.2015).
- Генератор текстов для КВН-www: *Генератор текстов для членов жюри КВН*. <http://kvnru.ru/txt.php> (30.5.2017).
- Генератор текстов-www: *Генератор текстов. Размножитель статей*. <http://linksfarm.ru/pages/226> (15.3.2017).
- Генератор тем-www: *Генератор тем рефератов, курсовых и дипломных работ по педагогике*. <http://ainte.ru/talkbox/ped/> (15.2.2015).
- Генератор хокку-www: *Генератор хокку онлайн*. <http://www.klikin.ru/hokku.html> (3.2.2015).
- Ген.Син Linksfarm-www: [*Генератор-синонимизатор*] Linksfarm. <http://linksfarm.ru/pages/226> (14.5.2017).
- Говорилка-www: *Говорилка*. <http://www.vector-ski.ru/vecs/govorilka/> (8.1.2017).
- Голосовой набор-www: *Блокнот для речевого ввода*. <https://speechpad.ru/> (17.5.2017).
- Искажение текста-www: *Искажение текста*. <http://pcvector.net/useful/214-generator-iskazheniya-teksta.html> (13.2.2017).
- Мотиватор-www: *Создаем мотиватор своими руками*. <http://timepost.ru/sozdaem-motivator-svoimi-rukami.html> (17.3.2017).
- Нонсенский генератор-www: *Нонсенский генератор стихотворной продукции*. <http://www.generator.nonsense.de/> (3.2.2015).
- Переворачивалка-www: *Переворачивалка символов*. <https://unicode-table.com/ru/tools/flip/> (13.2.2017).
- Рандомизатор текста-www: *Рандомизатор текста*. <http://miniwebtool.ru/> (8.1.2017).
- Рандомизатор текста-www2: *Рандомизатор текста* [2]. <http://xbb.uz/randomizator/> (13.2.2017).
- Рандомизатор-www: *Рандомизатор*. http://www.solarix.ru/for_developers/docs/randomizer.shtml (14.5.2017).
- Синонимайка-www: *Синонимайка*. <http://rabota-kopirait.com/index/sinonimajka/0-202> (6.5.2017).
- Синонимизатор Intop-www: *Генератор текстов [Синонимизатор Intop]*. http://intop24.ru/article_22.php (19.3.2017).
- Случайные данные-www: *Случайные данные для генерирования идей*. <http://www.toptrening.ru/articles/1023/> (20.4.2017).
- Сниппет-www: *Сниппет*. <https://ru.wikipedia.org/wiki/Сниппет>. 20.2.2017.
- Созд_свою_ист-www: *Генератор текста: Создай свою безумную историю*. <https://www.popmech.ru/editorial/248852-generator-teksta-sozday-svoyu-bezumnyu-istoriyu/> (20.3.2017).
- Стихоробот-www: *Стихоробот*. http://neogranka.ru/generator_stihov.html (1.6.2017).
- Стохастические стихотворения-www: *Стохастические ЭВМ стихотворения истина*. <https://katab.asia/2011/09/17/st0/> (20.3.2017).

- Текст-инверт-www: *Текст-инверт*. <http://textinvert.ru/> (4.3.2017).
- Чтение по складам-www: *Чтение по складам*. <http://www.abakbot.ru/online-5/128-sklad> (18.5.2017).
- Яндекс Рефераты-www: *Яндекс Рефераты*. <https://yandex.ru/referats/> (8.1.2017).
- Яндекс Переводчик-www: *Яндекс.Переводчик*. <https://translate.yandex.ru/> (15.10.2017).
- Blindtextgenerator-www: *Blindtextgenerator*. <http://www.blindtextgenerator.de/> (1.6.2017).
- Delirium-www: *Delirium 1.8*. <http://ppc-seo.blogspot.com/2007/10/delirium-18.html> (20.3.2017).
- Fake News-www1: *Fake News*. <https://twitter.com/fakenewsgen?lang=de> (15.3.2017).
- Google Übersetzer-www: *Google Übersetzer*. <https://translate.google.com/> (15.10.2017).
- Litlipsum-www: *Litlipsum*. <https://www.macwelt.de/downloads/Erstellt-Blindtext-fuer-Layouts-hat-aber-klare-Schwaechen-Blindtext-erstellen-mit-Little-Ipsum-1-1-1-3223280.html> (20.5.2017).
- Lorem Ipsum-www: *Lorem Ipsom*. <http://www.lipsum.com/> (15.3.2017).
- Markov Model Tweets-www: *Markov Model Generated Tweets*. <https://tweet-generator-alex.herokuapp.com/> (20.3.2017).
- Neuromzan-www: *Неуромзан*. <https://twitter.com/neuromzan> (12.2.2017).
- Randomtextgenerator-www: *Randomtextgenerator*. <http://randomtextgenerator.com/> (20.5.2017).
- SCIgen-www: *SCIgen*. <https://ru.wikipedia.org/wiki/SCIgen> (1.6.2017).
- Seot-www: *Бесплатный онлайн-генератор ключевых слов с текста*. <http://seotool.by/analiz/seo/keywordstext.php> (16.5.2017).
- SeoОптимизатор-www: *Seo Оптимизатор текста*. <http://www.seotxt.com/service/optimizer/> (16.5.2017).
- Translate.Eu-www: *Translate.eu*. <http://translate.eu/german/german-russian/> (15.10.2017).
- Translate.Ru-www: *PROMT: Online-Translator*. <http://www.online-translator.com/?prmtlang=de> (15.10.2017).

Литература

- Авт породж текста-www: *Автоматическое порождение текста*, <http://tp1-it.wikispaces.comАвтоматическое+порождение+текста> (20.3.2017).
- Джордж, Ф. 1963. *Мозг как вычислительная машина*, Москва: Изд-во иностранной лит-ры [= George. F. H. *The Brain as a Computer*, 1961].
- Добрушин, Р. Л. (отв. ред.) 1989. *Лингвистическое обеспечение системы Этан-2*, Москва: Наука. [Авторы Апресян, Ю. Д. / Богуславский, И. М., / Иомдин, Л. Л. / Лазурский, А. В. / Перцов, Н. В. / Санников, В. З. / Цинман, Л. Л.]
- Коваль, С. А. 2005. *Лингвистические проблемы компьютерной лингвистики*, Санкт-Петербург: Изд-во С.-Петербургского ун-та.

- Леонтьев, А. А. (отв. ред.) 1974. *Основы теории речевой деятельности*, Москва: Наука.
- Леонтьев, А. А. 2003². *Слово в речевой деятельности: Некоторые проблемы общей теории речевой деятельности*, Москва: Едиториал УРСС.
- Марковские цепи-www: *Генераторы псевдоосмысленных текстов на основе цепей Маркова*, <https://www.cy-pr.com/wiki/Марковские+цепи> (20.3.2017).
- Налимов, В. В. 2003³. *Вероятностная модель языка: О соотношении естественных и искусственных языков*, Томск / Москва: Водолей Publishers.
- Павлов, А. С. / Добров, Б. В. 2009. Методы обнаружения поискового спама, порожденного с помощью цепей Маркова, in: Калининченко, Л. А. (отв. ред.) *Труды 11^й Всесоюзной научной конференции „Электронные библиотеки: перспективные методы в технологии, электронные коллекции“ – ESDL '2009*, Петрозаводск: КарНЦРАН, 311–317.
- Пиотровский, Р. Г. 1979. *Инженерная лингвистика и теория языка*, Ленинград: Наука.
- Раскин, В. В. 1971. *К теории языковых подсистем*, Москва: Изд-во Московского ун-та.
- Ревзин, И. И. 1967. *Метод моделирования и типология славянских языков*, Москва: Наука.
- Соболева, П. А. 1972. Моделирование словообразования, in: Шаумян, С. К. *Проблемы структурной лингвистики 1971*, Москва: Наука, 165-212.
- Сорокин, В. Н. 1992. *Синтез речи*, Москва: Наука.
- Стемминг-www: *Стемминг*. <https://ru.wikipedia.org/wiki/Стемминг> (14.5.2017).
- Степанов, Г. В.; Панфилов, В. З. (отв. ред.) 1983. *Онтология языка как общественного явления*, Москва: Наука.