Инженерная лингвистика в контексте современной “Информации 4.0”

Language Engineering in the Framework of Modern “Information 4.0”

Л.Н. Беляева[1] **С.И. Богданов**[1] **Т. Горностай**[2]
Larisa Beliaeva[1] **Sergey Bogdanov**[1] **Tatiana Gornostay**[2]
lauranbel@gmail.com *rector@herzen.spb.ru* *gornostaja@tilde.com*

[1] Российский государственный педагогический университет им. А. И. Герцена,
Санкт-Петербург, Российская Федерация
[2] Тилде, Рига, Латвия

[1] Herzen State Pedagogical University of Russia,
Saint Petersburg, Russian Federation
[2] Tilde Company, Riga, Latvia

Abstract

Modern state of technology and science is defined by the potential of industrial automation processes (Industry 4.0) and appropriate presentation of information on the project under development and implementation. This potential is still determined by methods and principles of engineering linguistics. The paper considers the competences a linguist should have in this new situation and the necessity of special training courses.

Keywords: *Language engineering, language technologies, Information 4.0, competences*

Аннотация

Современное состояние технологии и науки определяются потенциалом процессов автоматизации в промышленности (Промышленность 4.0) и соответствующими способами представления информации к разрабатываемым проектам (Информация 4.0). Этот потенциал по-прежнему определяется принципами и методами инженерной лингвистики. В статье рассматриваются компетенции лингвиста в этой новой ситуации и специфика разработки специальных учебных курсов.

Ключевые слова:*Инженерная лингвистика, лингвистические технологии, Информация 4.0, компетенции специалиста*

1 Введение

Термин *инженерная лингвистика* был введен Раймондом Генриховичем Пиотровским еще в середине 60-х годов прошлого века. Под инженерной лингвистикой им понималось инженерное моделирование различных видов языковой компетенции – лингвистические технологии, предполагающие компьютерную реализацию разрабатываемых моделей. Подробное описание этой отрасли знаний и сути термина первым опубликовал Александр Михайлович Кондратов [Кондратов 1966], блестящий популяризатор науки, работу которого сам Р.Г. Пиотровский оценил очень высоко. Теоретическое осмысление этого направления можно найти в монографии Раймонда Генриховича [Пиотровский 1979]. В зарубежной лингвистике термину *инженерная лингвистика*по объему понятия практически соответствует термин *language engineering*. Современные подходы к решению задач в этой области с одной стороны развивают заложенные ранее принципы анализа и обработки текстов на естественном языке, с другой – модифицируются на основе новых технологий и новых требований к результатам создания и анализа текстов в различных гуманитарных и технических системах.

Лингвистические технологии, разрабатываемые в рамках инженерной лингвистики, охватывали все направления исследования текстов, которые сам Р.Г. Пиотровский объединял в своей концепции лингвистического автомата [Беляева, Пиотровский 2012]. И в этой концепции можно выделить два основных направления: вероятностное моделирование и исследование возможностей его применения для решения различных задач автоматической переработки текста, и разработку лингвистических технологий, эту переработку обеспечивающих. Сегодня оба эти направления приобрели особую важность для решения задач, относящихся к прикладной филологии в целом.

2 Основные направления инженерной лингвистики в системе современных технологий

Лингвистические технологии, разрабатываемые в рамках инженерной лингвистики, охватывали все направления исследования текстов, которые сам Р.Г. Пиотровский объединял в своей концепции лингвистического автомата [Беляева, Пиотровский 2012]. В этой концепции можно выделить два основных направления:

- вероятностное моделирование и исследование возможностей его применения для решения различных задач автоматической переработки текста,
- разработку лингвистических технологий, эту переработку обеспечивающих.

Сегодня оба эти направления приобрели особую важность для решения задач, относящихся к прикладной филологии в целом.

Если рассматривать проблемы вероятностного моделирования, то сегодня появление современных вычислительных систем, мощность которых, как известно, каждый год удваивается, особым образом повлияло на развитие инженерной лингвистики в области применения вероятностных и статистических подходов. Применение современных компьютеров и столь же современной периферии определило возможность вычисления сложных оценок поведения слова в тексте с очень большой скоростью, а также реальность сохранения огромных лингвистических данных (*Big Data*). Этим

во многом определяется новые подходы к смысловому анализу текста (sentiment analysis) на основе таких моделей поведения отдельных слов и пар слов в тексте как латентное размещение Дирихле (*Latent Дирихле Allocation – LDA*), модель фон Мизеса-Фишера (*von-Mises Fisher – vMF*), дискриминативная вероятностная модель (*Discriminative Probabilistic Model – DPM*) и др. [He et al. 2009]. В то же время следует учитывать, что большинство применяемых вероятностных оценок и статистических метрик по сути являются эвристиками, их адекватный выбор требует не только математического, но и лингвистического осмысления и обоснования. Так, например, переход от представления документа или слова как точки в пространстве (вероятно, текстов) к векторному представлению требует дополнительного рассмотрения (ср., например [Морозова 2013]) и лингвистического доказательства.

Если говорить о современном развитии лингвистических технологий, то следует учитывать, что сегодня развитие науки и техники во многом определяется степенью внедрения информационных технологий при реализации новых научных проектов и/или при разработке и внедрении конкретной научной и/или технической продукции. Недавно введенный термин Промышленность 4.0 (*Industry 4.0*) относится к современному подходу к автоматизации и обмену информацией в промышленном производстве [Gollner 2016]. Особенностью этого подхода является достижение максимальной гибкости производственных процессов за счет передачи оборудованию все большего числа распределенных вычислений и независимых решений, принимаемых на основе цифровой информации.

Естественно, что уровень реализации принципов и методов Промышленности 4.0 зависит от того, насколько стандартизированы методы создания, обмена и использования информации о разрабатываемом проекте, производстве, об эксплуатации конкретного технического устройства и о материальном обеспечении. Подобная информация создается в виде текстов на естественном языке – технической документации на всех этапах реализации проектов, от Технического Задания до рекламного проспекта, от инструкции по эксплуатации до руководств пользователя. От качества этих документов, создаваемых на исходном естественном языке и затем переводимых на все языки распространения продукции, зависит возможность применения высоких уровней автоматизации при их интерпретации и публикации. В контексте Промышленности 4.0 определяются следующие важные характеристик Информации 4.0:

- **молекулярность** – нет отдельных документов, формируются информационные молекулы, которые в дальнейшем могут соединяться в тексты в зависимости от контекста и целей использования,

- **динамичность** – непрерывность обновления и модификации молекул информации и текстов в целом,

- **свобода выбора пользователем** – информация предлагается, а не поставляется вместе с продуктом,

- **глобальность** – возможность доступа к информации через Интернет из любой точки мира, интерактивная, доступная и удобная для поиска,

- **спонтанность** – возможность свободного определения в зависимости от контекста, цели и ситуации использования,

- **профилированность** – автоматизация создания текстов определенной структуры и лексического состава [Gallon, 2016].

Следовательно, информация, представленная на естественном языке (как пра-

вило, на языке контролируемом) в виде научной и/или технической документации, должна быть подготовлена для использования в различных ситуациях, должна быть сформулирована так, чтобы обеспечить возможность ее динамичного приспособления к различным сценариям производства, эксплуатации и материального обеспечения. Информация должна быть структурирована и сформирована так, чтобы ею можно было обмениваться на любых этапах реализации проекта. Именно здесь и возникает необходимость использования методов инженерной лингвистики в ее сегодняшнем представлении.

Сама структура доступа к информации и способы ее использования с новыми интерфейсами и революционными подходами к информации, далеко ушедшими от традиционного представления текста, активно меняется. Современные средства работы с информацией (*toolkits*) должны в будущем объединить:

• самодокументирующие устройства (*self-documented devices*), позволяющие извлекать из текста и формировать прогностические и контекстные указания,

• дополненную реальность (*Augmented Reality*), уже имеющуюся на планшетах и мобильных устройствах,

• встроенные инструментальные средства типа очков с искусственным интеллектом (*smart glasses*).

Особое значение обмен информацией и данными приобретает в рамках так называемого Интернета вещей (*Internet of Things - IoT*), при организации которого происходит обмен не просто информацией об объектах, а самими объектами. Лингвистические и технологические проблемы, связанные с новыми формами и методами представления информации, обсуждались на очередной конференции *tcworld*, происходившей в 2016 г. в Штутгарте. В рамках этой конференции рассматривались проблемы выбора инструментальных средств и подходов к новым технологиям, включая разработку новых учебных программ, позволяющих подготовить переводчика, способного решать новые задачи работы с информацией на естественном языке.

Таким образом, для активного развития науки и техники необходима информация, фиксируемая в текстах технической документации, которая может сопутствовать всему жизненному циклу научной и/или технической продукции и использоваться самыми разными способами. Такая информация, способная на поддержку киберфизических систем Промышленности 4.0, называется Информацией 4.0 (*Information 4.0*) и создается с помощью специализированных систем создания текстов с опорой на информационные технологии. Сегодня наиболее активно используемой и столь же активно обсуждаемой специализированной системой является DITA (*Darwin Information Typing Architecture*), базовая спецификация которой определяет набор типов документов, предназначенных для создания документов авторами и организации тематически-ориентированной информации, а также и набор механизмов для объединения, распространения и ограничения типов документов [DITA Forum 2016]. Система скачивается бесплатно и позволяет решать различные задачи по созданию и форматированию текстовых документов.

В основе представления научной и технической документации в рамках подхода Информация 4.0 лежит понятие авторской разработки структурированного контента (*structured content authoring*), которая состоит в разбивке содержания на небольшие части, называемые тематическими разделами (*topics*), которые впоследствии собираются с помощью карт (*maps*) для того, чтобы создать окончательный вариант контента. Этот подход отличается от общепринятого варианта создания неструктури-

рованных документов с использованием инструментальных средств подготовки текстов. Инструментальные средства разрабатывались и применялись для того, чтобы оптимизировать продуцирование и поддержание больших массивов текстовых документов на основе систем, которые позволяют создавать тексты параллельно, избегая дублирования контента за счет повторяющихся тематических разделов. Тем самым облегчается модификация текстов, связанная с разработкой новых версий изделия, уменьшаются расходы на услуги переводчиков и т.д.

В основе нового подхода лежит анализ продуктивности (*productivist approach*), при котором степень детализации конкретных тематических разделов определяется задачами создания научной и технической документации и потенциально отделена от самого содержания, т.е. от тех тем, которые реально обсуждаются в тексте [Lacroix 2016].

3 Подготовка современных специалистов в области работы с информацией

Специалистам, работающим с новыми формами представления информации, соответствует английский термин *language worker*, который можно приблизительно перевести как специалист в области переработки текстов [Беляева 2016]. Такой термин используется как объединяющая номинация для терминологов, переводчиков, для всех тех, кто создает техническую документацию (технических писателей – *technical authors*, *technical writers*), специалистов по передаче технической информации (*technical communicators*), компьютерных лингвистов и т.д.

Сегодня и обработка текста на естественном языке, а также научный и особенно технический перевод включены в единый технологический процесс, осуществляемый по заранее определенным правилам, в соответствии с графиком выполнения работы и международными стандартами. Уровень развития лингвистических технологий определяет необходимость уточнения места и функций технического перевода и самого технического переводчика в особой технологической цепочке, включающей использование систем машинного перевода, комплекса автоматизированных словарей, предметно ориентированного корпуса текстов, комплекса прикладных программ [Беляева 2016].

Поскольку умение перевести специальный текст вырабатывается тогда, когда человек способен создать этот текст на родном языке, то профессиональные переводчики, терминологи, технические писатели должны обладать базовыми компетенциями в области создания специальных текстов на родных и иностранных языках, а также в области их перевода и обработки. В качестве такой обработки может рассматриваться извлечение информации, а также создание вторичных текстов любого типа и назначения.

Выполнение всех этих видов работ требует от специалистов в области обработки текстов

1) знания типологии специальных и технических текстов на родном (русском) языке и иностранных языках, их различий и особенностей;

2) умения создавать все типы специальных текстов на родном языке и иностранном языке;

3) умения переводить тексты с учетом различий в требованиях к специальным

текстам в различных культурах.

К сожалению, в нашей стране специалистов в области разработки технической документации не готовят. Необходимые сегодня специалисты должны обладать рядом стандартных компетенций в области планирования своей работы, создания специального текста, учитывая такие требования как ясность, краткость, простота выбираемых выражений, использование корректной терминологии, активного залога, полных синтаксических конструкций, отказ от использования синонимических терминов; анализа и редактирования получаемого результата.

Однако Информация 4.0 требует и совершенно новых компетенций, к которым в рамках инженерной лингвистики относятся:

• способность собирать, анализировать и отбирать подходящую информацию, чтобы разрабатывать информационный продукт,

• способность выбирать стратегию разработки продукта для того, чтобы создавать соответствующие информационные продукты для различных целей и потребителей,

• способность гарантировать, что информация является извлекаемой и доступной, представляет связную ментальную модель и согласуется по продуктам и средам

• умение выбирать аппаратные средства и программное обеспечение,

• достаточное понимание предметных областей, которые являются релевантными для специалистов по распространению технической информации (информатика, машиностроение, физика и т.д.), чтобы быть способными сотрудничать с экспертами в предметной области,

• знание основных принципов и методов терминоведения,

• способность формировать ресурсные и лексикографические базы данных и корпуса текстов для решения профессиональных задач [ср. Meex, Karreman 2016].

Две последние компетенции относятся к работе с терминологией, поскольку в новой информационной среде технический писатель, менеджер по продукции и терминолог выявляют новую терминологию, которая появляется по мере разработки продукции, в результате ее сертификации и документирования. При этом учитываются все виды документации: описания и спецификации, руководства пользователя и отчеты, пользовательские интерфейсы, сообщения об ошибках и системные сообщения и т.п., а также создаются словари, использование которых является обязательным.

4 Выводы

К сожалению, приходится констатировать, что подготовка технических писателей далека от требований, которые предъявляются к ним новыми формами представления информации и работы с ней. Все сказанное выше позволяет утверждать необходимость введения специальной подготовки специалистов в области переработки текстов, определяющей развитие специальных профессиональных компетенций в работе с Информацией 4.0 и использования специальных информационных технологий создания технической документации. Кроме того, следует подчеркнуть, что современные специалисты должны учитывать принятое «разделение труда», заключающееся в том, что, например, терминолог, переводчик и специалист по рекламе имеют различный функционал, но все они должны уметь работать в команде.

Использование лингвистических технологий и конкретных систем подготовки информации давно стало элементом профессиональной работы переводчика и терминолога, а для специалиста - средством извлечения знаний из текста. Грамотное использование ресурсов лингвистических технологий: электронных баз данных и знаний, систем машинного перевода, тезаурусов, онтологий, систем проверки орфографии, систем доступа к информации по различным сетям передачи данных давно вошло в реальный обиход специалистов в различных областях знаний. Современный специалист работает сегодня в высокотехнологичной среде и имеет возможность выбора удобной для него конкретной информационной системы. Поэтому собственную ресурсную базу любой профессионал должен научиться компоновать из различных систем обработки информации, уметь подбирать автоматизированные словари в соответствии со своими запросами и сферой деятельности, знать их ограничения и возможности, знать, какие лексикографические источники отсутствуют в электронном формате. И специалист, и терминолог, и переводчик должны хорошо представлять себе ресурсы Интернета и требования, предъявляемые сегодня к тому, что называется Информация 4.0.

Таким образом, можно утверждать, что с развитием компьютерной техники и технологий ее использования инженерная лингвистика как метод работы с текстом на естественном языке обретает новое и важное звучание.

Список литературы

[Beliaeva 2015] Beliaeva, L.N. Lingvisticheskie tekhnologii v sovremennom setevom prostranstve: language worker v industrii lokalizacii [Linguistic technologies in modern network space: language worker in localization industry]. Sankt Peterburg: Knizhnyj dom, 2016. – 134 s. (In Russian) = Беляева Л.Н. Лингвистические технологии в современном сетевом пространстве: language worker в индустрии локализации. СПб.: Книжный дом, 2016. – 134 с.

[Belyaeva, Piotrovskij 2012] Belyaeva L.N., Piotrovskij R.G. (2012) Inzhenernaya lingvistika v Gercenovskom universitete: teoriya inzhenerno-lingvisticheskih issledovanij i praktika razrabotki informacionnyh system [Language Engineering in Herzen University: theory of Language Engineering research and informational systems building practice] // Nauchnoe mnenie, [Scientific opinion] № 9. SPb. S.37-45 (In Russian) = Беляева Л.Н., Пиотровский Р.Г. Инженерная лингвистика в Герценовском университете: теория инженерно-лингвистических исследований и практика разработки информационных систем // Научное мнение, № 9. СПб, 2012. С. 37-45

[DITTA FORUM 2016] DITA Forum // Towards a European Competence Framework // tekom-Jahrestagungund tcworld conference in Stuttgart. Zusammenfassungen der Referate – Stuttgart: tcworld GmbHVerantwortlich, 2016. Pp. 51-61

[He et al. 2013] He Q., Chang K., Lim E., Banerjee A. (2013) Keep It Simple with Time: A Re-examination of Probabilistic Topic Detection Models. Retrieved 01.10.2017 from http://wwwusers.cs.umn.edu/ banerjee/papers/09/pami-TD t.pdf

[Gallon 2016] Gallon R. Information 4.0, the Next Steps //Towards a European Competence Framework // tekom-Jahrestagungund tcworld conference in Stuttgart. Zusammenfassungen der Referate . Stuttgart: tcworld GmbHVerantwortlich, 2016. Pp. 95-97

[Gollner et al. 2016] Gollner J. Information 4.0 for Industry 4.0 // Towards a European Competence Framework // tekom-Jahrestagungund tcworld conference in Stuttgart. Zusammenfassungen der Referate – Stuttgart: tcworld GmbHVerantwortlich, 2016. Pp. 93-94

[Kondratov 1966] Kondratov A.M. Zvuki i znaki. [Sounds and Symbols] M.: Znanie, 1966. – 207 s. (In Russian) = Кондратов А.М. Звуки и знаки. М.: Знание, 1966. – 207 с.

[Lacroix et al. 2016] Lacroix F. Writing for the 21st Century // Towards a European Competence Framework // tekom-Jahrestagungund tcworld conference in Stuttgart. Zusammenfassungen der Referate – Stuttgart: tcworld GmbHVerantwortlich, 2016. Pp. 102-106

[Meex et al. 2016] Meex B., Karreman J. TecCOMFrame. Towards a European Competence Framework // Towards a European Competence Framework // tekom-Jahrestagungund tcworld conference in Stuttgart. Zusammenfassungen der Referate – Stuttgart: tcworld GmbHVerantwortlich, 2016. Pp. 486-489

[Morozova 2013] Morozova Yu.I. Postroenie semanticheskih vektornyh prostranstv razlichnyh predmetnyh oblastej [Building semantic vectorspace for different subject fields] // Informatika i ee primenenie [Informatics and its application], 2013. Vol. 7, Issue 1. S. 90-93 (In Russian) = Морозова Ю.И. Построение семантических векторных пространств различных предметных областей // Информатика и ее применение, 2013. Т. 7, Вып. 1. С. 90-93

[Piotrovskij 1979] Piotrovskij R.G. Inzhenernaya lingvistika i teoriya yazyka . [Language Engineering and Language Theory]. L.: Nauka, 1979. – 112 s. (In Russian) = Пиотровский Р.Г. Инженерная лингвистика и теория языка. Л.:Наука, 1979. – 112 с.