МАШИННЫЙ ФОНД РУССКОГО ЯЗЫКА: ВНЕШНЯЯ ПОСТАНОВКА

Осенью 1978 г. один специалист по программированию, имевший опыт создания диалоговых систем общения с ЭВМ, получил приглашение выступить с докладом перед сообществом физиков, биологов и программистов, собравшихся на первую (но ставшую впоследствии традиционной) конференцию Диалог-78 в Биологическом центре АН СССР в Пущине-на-Оке. Доклад содержал изложение общей гипотезы о продуктивности диалога человека с ЭВМ на естественном языке при условии ограничения этого диалога так называемой "деловой прозой". Надежда на продуктивность гипотезы базировалась на другой гипотезе о внутренней формализованности любого устойчивого процесса производства и возникающих в нем человеческих отношений. Эта тема сама по себе интересна, в том числе и для обсуждения на этом совещании, но дело не в основной линии доклада. Говоря об общенаучных предпосылках диалога с ЭВМ на естественном языке, докладчик, между прочим, сказал:

"Любой прогресс в области построения моделей и алгоритмов останется, однако, академическим упражнением, если не будет решена наиважнейшая задача создания Машинного фонда русского языка. Это фундаментальная проблема, решение которой будет иметь очень большую научную, общекультурную и прикладную ценность. Не мне, конечно, составлять спецификацию такого фонда, но думается, что по крайней мере он должен содержать полный словарь и генератор словоформ, а также формализованный толковый словарь (тезаурус) русского языка. Очень хотелось бы видеть, что создание машинного фонда русского языка квалифицированными лингвистами опережало бы создание производственных лингвистических систем, потому что это не только бы позволило избежать дублирования больших усилий, но и защитило бы здоровую ткань русского языка от самоуправства и неквалифицированного подхода".

Это была первая высказанная вслух мысль о Машинном фонде русского языка. Тот факт, что я же и был докладчиком, не имеет особого значения; сейчас мне хочется обратить внимание на то, что эта постановка вопроса была сделана не лингвистом и адресована нелингвистической аудитории.

Надо сказать, что на этом же докладе мне был задан вопрос, заинтересовались ли лингвисты идеей Фонда. В тот момент вопрос был явно преждевременен, однако сегодня реакция превзошла всякие ожидания, и вообще это движение набирает силу столь стремительно, что полагать его реакцией на чей бы то ни было конкретный призыв было бы просто смешно.

И все же, оставаясь профессионально вне лингвистики, я как специалист по информатике беру на себя смелость предложить этот внешний разговор о Машинном фонде русского языка как о важнейшей задаче, решение которой очень сильно повлияет как на саму лингвистику, так и на ее место в обществе. По существу, должна произойти некоторая смена парадигмы этой науки и, собственно, это-то и придает мне смелости, так как известно, что предпосылки к смене парадигмы созревают зачастую вне рассматриваемой науки. Кроме того, я беру в союзники тридцатилетний опыт наблюдения за компьютеризацией самых разных областей деятельности, в том числе и научной, и, если угодно, мой собственный полувековой опыт применения русского языка: ведь человек, который пишет, читает, говорит, думает, вообще живет в своей родной языковой среде, относится к ней очень глубоко и активно. Кажется, у нас нет формально объявленного Общества любителей русской словесности, но на самом деле такое сообщество существует, объединяя сотни миллионов человек. От этого сообщества науке о языке передаются мощные импульсы потребностей и интереса, удовлетворению которых Машинный фонд может сильно помочь. В глазах этих людей русистика – это спящий Гулливер, которому пора проснуться и заговорить о себе полным голосом. Вычислительная техника может для этого оказаться подходящим будильником.

Каждый знает, какую роль сыграла в становлении В. И. Ленина как вождя русской революции написанная им в молодости работа "О развитии капитализма в России". Но сделаем еще один шаг. Эта работа не могла бы появиться без так называемой земской статистики. Я сошлюсь на свидетельство академика А. Г. Аганбегяна, который в одном из своих выступлений говорил об общей значимости земской статистики в России во второй половине XIX в., ознаменовавшей в экономической науке того времени переход от наблюдательного изучения к измерительному. Действительно, только такой переход позволил выдвинуть экономические методы управления, создать экономические модели – в целом приблизить экономику к разряду точных наук, сделать ее стратегическим оружием в развитии общества.

Между экономикой и языком как системами существует немало сходных черт. И та и другая являют как свою изменчивость, так и свою устойчивость в миллиардах единичных актов человеческой деятельности. В каждой из этих систем их объективные законы получаются интегрированием индивидуальных актов сознания, которые при всей своей независимости, поддерживаемой волей участника событий, хотят тем не менее явно опираться на эти законы. Так вот, если поддерживать эту аналогию, то можно сказать, что

русистике как науке тоже нужно в полной мере перейти от наблюдательного периода к измерительному. Конечно, этот переход давно наметился, но появление вычислительной техники дает этой проблеме совершенно новую постановку. И еще раз приведенная аналогия с экономикой показывает русистике как диапазон возможностей, так и величину дистанции, которую надо преодолеть, чтобы получать от языка прямые ответы на сто тысяч вопросов, которые ему хочется задать.

Итак, Машинный фонд русского языка может стать могущественным концентратором нашего знания о языке, знания полного, не отстающего от хода времени, детального и обобщенного вместе, подвижного и инерционного, накапливающего все предыдущие временные срезы.

Еще один аспект отношения "ЭВМ в языке" – это проблема овладения языком. Признано, что сейчас в культуре налицо две разнонаправленные тенденции. Одна из них – это явное увеличение роли печатного, услышанного и произнесенного слова. Круг словесного общения сейчас очень интенсивен. Средства массовой информации многократно повышают интенсивность такого общения, хотя, возможно, и делают его не столь прямым. Феноменальный рост личных книжных собраний дополняет эту интенсификацию. В то же время отмечается некоторое снижение уровня грамотности, не очень хорошо обстоит дело с изучением русского языка в национальных школах.

Можно, по-видимому, найти этому разумное объяснение путем более внимательного изучения обстановки в школе и в семье, пытаться улучшать положение традиционными методами. Есть, однако, ряд свидетельств тому, что широкое внедрение учебных ЭВМ в школе позволит радикально изменить ситуацию с изучением русского языка. Первые опыты применения ЭВМ в школе показывают, что машина может стать активным партнером в учебном процессе и не менее стимулировать самого ребенка, создавая, где нужно, игровую обстановку, увеличивая элемент состязательности, обеспечивая индивидуальность и мгновенность реакции, побуждая к сознательному знанию и одновременно стимулируя и тренируя всяческую моторику, закрепляющую добытое знание.

Другой аспект интенсификации работы с русским языком я бы назвал проблемой "экспансии" языка, имея в ввиду его роль как языка межнационального общения и как мирового языка. И здесь я не вижу никаких причин тому, чтобы уклоняться от культурного соревнования с другими мировыми языками, и прежде всего с английским. И как раз тут у нас не все в порядке. Зайдите в магазин иностранной литературы и посмотрите на обилие словарей английского языка: и один толковый словарь, и другой, и английский, и американский, словари синонимов, жаргонов, словари для студентов и школьников, с картинками и без, словари рифм, обратные словари и многое-многое другое. Жаль, что мы не обнаруживаем у себя такого обилия, такой кажущейся избыточности, такого языкового напора.

Конечно, мы делаем очень много. У нас есть замечательные словари. Например, у нас вышел великолепный однотомный "Энциклопедический словарь". Но ведь его нужно издавать каждый год, поддерживая конъюнктурные изменения. Праздником русской культуры стало появление Лермонтовской энциклопедии, но где энциклопедия Пушкина, Достоевского, Толстого, Чехова, Горького? Почему Академический словарь русского языка должен выходить один раз в пятнадцать – двадцать лет? Можно этому найти много причин и обнаружить их объективность. Но дело каждого преодолевать те причины, которые зависят от него.

Так вот, одна из причин – это та, что словарное и вообще лексикографическое дело для ведущих языков – английского, французского, немецкого – уже давно поставлено на машину и соединено с достижениями полиграфии, прежде всего с фото- и электронным набором. Можно и должно дорожить накопленными фондами в их теперешнем виде, уважать сложившиеся методы работы, но там, где идет речь об освоении пространства, старая карета и новая машина – не конкуренты.

Мне эта задача представляется очень важной. Она важна не только для повышения культурного уровня нашего общества и не только для усиления интегративных процессов в нашей стране, но и для реализации функции русского языка как языка международного общения. А для того чтобы понять, что тут все не так просто, достаточно просмотреть международные научные журналы, которые публикуют статьи как на русском, так и на английском языках, и посмотреть, как изменилось соотношение языков опубликованных текстов за последние двадцать лет.

Продолжая линию внешних постановок, я хотел бы обернуть отношение и поговорить о "языке в ЭВМ". Сначала небольшое философское отступление. За последние десятилетия начинает складываться представление о трех формах существования материи, энергии, отражающей структуру материи, и сам процесс ее познания. Мало того, три ипостаси дают определенную периодизацию развития человеческой цивилизации: освоение вещества, создание материального производства, овладение принципом единства материи (от глубины тысячелетий до начала XX в.); затем освоение энергии, создание энергетического изобилия, овладение принципом единства энергии и ее связи с веществом (от XVIII до XXI в. в надежде на

овладение термоядерной энергией) и, наконец, освоение информации, создание полной информационной доступности, овладение принципом единства процессов обработки информации в естественных и искусственных системах (от XX в. в будущее).

Здесь нам через развитие средств электросвязи, вычислительных средств, автоматики и предстоит где-то в середине XXI в. войти в период "сплошной информатизации", который в своем внешнем выражении будет характеризоваться следующими показателями:

– грандиозная сеть электронной связи с полумиллиардом входов;
– полная автоматизация техносферы с парой миллиардов встроенных микроЭВМ;
– порядка двухсот-трехсот миллионов персональных ЭВМ и интеллектуальных терминалов, подключенных к сети связи;
– десятимиллионная иерархия универсальных ЭВМ, поддерживающая управление обществом и сетью связи;
– перенос на машинные носители практически всей информации, циркулирующей в обществе.

Все это должно не только существовать, но и воплощать полноту и доступность знания, хранить в себе достоверную информационную модель мира, быть в постоянном употреблении всеми людьми, реализуя на новой основе принцип единства человеческого рода. Возникает вопрос: как человек будет общаться с этой инфосферой, как он будет побуждать машины к действию, как он будет черпать из этого грандиозного фонда знания, как он будет относиться к новому жителю своего дома – компьютеру?

При первых же размышлениях над этим вопросом сразу возникает уже достаточно хорошо известная номенклатура конкретных технических проблем: организация диалога с базами данных; составление баз знаний и общение с ними (база знаний отличается от базы данных, грубо говоря, тем, что представляет теорию более высокого порядка: если в базе данных – конкретные факты, то в базе знаний – общие суждения и определенные дедуктивные свойства); извлечение смысла из текста в виде команд, фактов, энциклопедических знаний и т. д.; пополнения знаний (очень деликатная задача, так как речь идет не о механическом добавлении нового факта или суждения, а об установлении всех возможных связей с уже накопленной системой знаний); машинный перевод; синтез текста. На последней проблеме я хотел бы остановиться более подробно.

Каждый из нас сейчас читает много текстов, которые почти исключительно написаны людьми. Мы должны готовиться к тому, что это человеческое начало в текстовой информации начнет уступать заметное место машине. Хочу предупредить, что я совершенно не посягаю на фантастические сюжеты машинного стихосложения или любовных писем робота. Я говорю о реальных, прозаических вещах, связанных прежде всего с прозой. Сейчас, например, автоматизация проектирования становится стержнем технического прогресса. но неотъемлемой частью автоматического проектирования становится синтез технических описаний. Причем это не отдельные фразы или перечни слов в каких-то фиксированных таблицах. Это нормальный связный текст. Уклониться от рассмотрения такой проблемы нет никакой возможности. Во все большем количестве конструкторских бюро изготовление технической документации – узкое место. Профессия "технического писателя" – самая дефицитная. Нет людей, которые имеют вкус к такой работе. Устранять человеческий язык из технической документации никак нельзя. Нам с этой техносферой надо сосуществовать веками, и тем более нам нельзя терять наше человеческое. И если мы привыкли читать нормальный текст, мы предпочтем читать именно его, даже если он и синтезирован на машине.

Кроме технической документации к этой синтетической прозе нужно будет отнести большое количество сводок, отчетов и других текстовых свидетельств. Здесь – мой особый призыв к лингвистам. Очень не хотелось бы обрекать деловую прозу на посредственность. На первый взгляд, деловая проза бесконечно далека от интереснейших и красивейших задач внутреннего изучения русского языка. Есть ли у деловой прозы коммуникационная сверхфункция, есть ли у нее одухотворенность и чем она может быть красива – все это тоже, как мне кажется, достойнейшая задача большой науки о русском языке.

Мы хотим как можно глубже познать природу языка, и в частности русского. Одним из выражений этого познания должна стать модель русского языка. Это формальная система, которая должна быть адекватной и равнообъемной живому организму языка, но в то же время она должна быть анатомически отпрепарированной, разъятой, доступной для наблюдения, изучения и изменения. Я хочу сопоставить эту еще не существующую модель с другой моделью, созданной человечеством для изучения неживой природы, – математическим анализом, и даже с одной лишь его ветвью – дифференциальными уравнениями. Нет никаких сомнений в том, что по сравнению с русским языком это куда более скромная область. И в то же время, посмотрите, как огромен интеллектуальный багаж этой части науки. Сколько мы имеем разных руководств, теорий, сколько методов решений дифференциальных уравнений, как мощен поток литературы, как тщательно мы учим этому в университетах и технических вузах. Конечно, мы учим и русскому языку. Но если, действительно, сравнить размах области знаний, связанных с моделированием форм неживой

материи, с нынешним статусом и объемом науки о языке, то станет ясно, что мы только в самом начале пути. Мне бы не хотелось этим предостережением сбивать ноту высокого оптимизма, на которой началась наша дискуссия, но думается, что лишь сознание необходимости и неотвратимости и качественного перелома в науке о языке, сознание ее общенаучной и государственной важности смогут создать тот запал и энтузиазм, без которого проблему Машинного фонда русского языка не поднять.