

© 1993 г. ХАНТ ДЖ.

ВЫЧИСЛИТЕЛЬНАЯ ЛИНГВИСТИКА В ЛЕТНЕМ ИНСТИТУТЕ ЛИНГВИСТИКИ

О ЛЕТНЕМ ИНСТИТУТЕ ЛИНГВИСТИКИ

Летний Институт лингвистики (Summer Institute of Linguistics — SIL) создан в США в 1934 г. Со дня своего основания он занимается проблемами описания языков малочисленных народов и перевода Библии на эти языки.

В первые годы существования лингвистические курсы читались только в летние месяцы, благодаря чему Институт и получил свое название. Теперь лингвистические курсы часто читаются круглый год, и не только в США, но и в Канаде, Великобритании, Франции, Германии, Австралии, Новой Зеландии, Японии и в Сингапуре. В университетах США такие курсы считаются аспирантскими, в других странах — приравниваются к университетским. Более короткие курсы организуются по мере надобности во многих других странах.

В последнее время SIL главное внимание уделяет проблеме распространения и повышения грамотности. Это побудило правительства Южной Америки поддерживать организацию двуязычных школ для индейцев.

SIL является одной из крупнейших лингвистических организаций в современном мире. Он насчитывает 6000 сотрудников, половина из которых — лингвисты. Работы ведутся приблизительно по 850 языкам. Примерно по 300 языкам работы завершены.

SIL финансируется церквями или спонсорами, поддерживающими отдельных сотрудников, SIL не финансируется правительствами, хотя правительственные организации иногда оказывают через SIL помощь странам третьего мира в финансировании программ ликвидации неграмотности.

ЛИНГВИСТИЧЕСКИЕ АСПЕКТЫ ПОЛЕВЫХ ИССЛЕДОВАНИЙ

Общий подход SIL к полевым исследованиям состоит в том, что каждый проект имеет дело с конкретным языком. В проекте участвуют обычно два человека. Они должны жить среди носителей изучаемого языка, учиться говорить на этом языке, вовлекать его носителей в лингвистические и антропологические исследования и использовать результаты этих исследований на благо этого народа. SIL оказывает таким проектным группам консультационную помощь.

Прежде чем начать работу над проектом, проводят различные обследования с помощью анкетирования. Оно включает в себя лексическое анкетирование, социо-лингвистическое анкетирование и тестирование билингвизма.

Приступая к осуществлению проекта, группа поселяется в выбранном месте, обучается языку, собирает фонетические и антропологические данные. После сбора данных начинается их лингвистический анализ. Он включает в себя фонологический, морфологический, лексический, синтаксический, семантический анализ и анализ дискурса. Эти исследования завершаются публикациями материалов и разработкой, а также последующей публикацией учебных пособий —

букварей, хрестоматий, фольклорных сборников. Во многих случаях сами информанты или другие аборигены входят в проектную группу, иногда, как, например, в Гане, берут на себя административное руководство проектами. В других странах поощряется создание институтов, подобных SIL.

Хотя я выше и описал в целом типичное полевое исследование, такие проекты бывают весьма разнообразными, как по трудности осуществления, так и по количеству трудозатрат. Особо трудными бывают исследования, когда лингвист и информант не владеют ни одним из языков друг друга, как например, в джунглях Южной Америки. Но и такие проекты нередко заканчивались успешно.

Мои собственные интересы лежат в области изучения различных типов мышления людей в разных частях света. Например, народы, являющиеся охотниками и собирателями, развивают некоторые когнитивные умения в гораздо большей степени, чем мы на Западе, но, с другой стороны, они не обнаруживают некоторых когнитивных способностей, свойственных европейцам. Это неизбежно оказывает часто не осознаваемое нами влияние на методику полевых исследований.

Некоторые языковые сообщества, с которыми работает SIL, очень малочисленны, и могут насчитывать менее сотни членов. С другой стороны, полевые исследования проводятся также в странах, население которых исчисляется миллионами человек. Иногда анкеты показывают, что люди, имеющие одно и то же имя, например, благодаря одному и тому же предку, не понимают друг друга. Поэтому необходимы несколько полевых исследований. В других случаях анкеты свидетельствуют, что люди, говорящие на номинально различных языках, прекрасно понимают друг друга.

При таких больших различиях проектов необходима гибкость в их организации. Если SIL суждено сыграть существенную роль в России, то, вероятно, необходимо развивать новые подходы к реализации проектов.

Др. К. ПАЙК

Др. К. Пайк, хорошо известный в СССР, работал в SIL с момента его основания. Он занимался тогда языками Мексики и позднее стал профессором лингвистики в Мичиганском университете. К. Пайк разработал модель грамматики (тагмемику) и обобщенную концепцию эмического и этического уровней описания языков.

К. Пайк в течение нескольких лет был президентом SIL, и за работу в SIL ему была присуждена Нобелевская премия. Недавно он был избран почетным членом Международной Лингвистической Ассоциации содействия преподаванию иностранных языков, организованной в Москве в июне 1990 г.

РОЛЬ SIL В ЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЯХ

Видимо, наибольший вклад в изучение многочисленных языков, внесенный сотрудниками SIL, состоит в сборе и первичной обработке данных. Это дает возможность проверять новые лингвистические теории и модели на реальном языковом материале, способствует построению новых моделей грамматики.

На материале многочисленных лингвистических опросов, проведенных сотрудниками SIL, и комментариев к ним других лингвистов SIL издает труд под названием "Ethnologue: Languages of the world". Эта книга содержит детальные сведения о языках всего мира (численность и место обитания носителей, о звуковых и знаковых языках и др.) и регулярно дополняется, уточняется и переиздается.

Работа по сбору данных требует организации новых 2500—3000 полевых исследований, проведение которых, по всей видимости, приведет к созданию новых лингвистических организаций по всему миру.

Особый тип анкетирования, разрабатываемый SIL, представляют тесты на понимание, в частности тест, предложенный Ю. Касадом (Eugene Casad). Простые словарные анкеты часто не дают адекватного представления о том, правильно ли понимают друг друга носители различных языков или диалектов. Поэтому и понадобился новый тест.

Развиваемая лингвистами теория дискурса, охватывающая высшие уровни грамматической иерархии, родилась благодаря работе над переводами Библии. Переводчик имеет дело не только со словами или предложениями, но и с аргументами, историческими событиями, иносказаниями и пр. При этом перевод в идеале должен достигнуть того же воздействия на читателя, что и первоисточник. Теория дискурса стала очень ценным инструментом для этого.

Вслед за трудами Ю. Найды (Eugene Nida) и др. SIL сыграл заметную роль в развитии теории перевода на Западе. Большой вклад внесли в него Дж. Бикмен, Дж. Кэллоу и М. Ларсон (John Beekman, John Callow, Mildred Larson).

Сотрудник SIL С. Гудчински (Sarah Gudchinsky) разработала лингвистическую теорию обучения грамоте и подготовке букварей. Применение этой теории определяется языком и реакциями его носителей.

По просьбе правительств Южной Америки были организованы школы с преподаванием на двух языках, с тем чтобы помочь аборигенам овладеть доминантной культурой, от которой они почти полностью изолированы.

Б. Мэнн и Д. Вибер (Bill Mann, David Weber) разработали теорию пакета CADA (Computer assisted dialect adaptation).

Сотрудники SIL опубликовали и продолжают публиковать множество трудов. К сожалению, большое количество из огромного множества данных недоступны другим исследователям, но в последнее время предпринимаются усилия создать общедоступные научные архивы.

SIL И КОМПЬЮТЕРЫ

Возможности, предоставляемые компьютерами для обработки такого большого количества материала, добываемого в полевых исследованиях, всегда привлекали внимание лингвистов. Первоначально отпечатанные данные направляли на обработку в вычислительные центры развитых стран. Главным был центр в университете штата Оклахома, который изготавливал конкордансы по текстам на различных языках и готовил публикации на машине DEC PDP 8. Использование компьютеров в поле долгое время оставалось мечтой.

Во второй половине 70-х годов компьютеры уменьшились в размерах настолько, что стали походить на телевизоры, снабженные клавиатурой. Но они дорого стоили, потребляли много энергии, их память была не более 64К, они не имели графических возможностей и лингвистически ориентированного программного обеспечения. Используемые тогда принтеры были громоздки и медленны. Но несмотря на это некоторые лингвисты-полевики покупали такие компьютеры и использовали их в поле. Постепенно было разработано нужное программное обеспечение, хотя и возникали трудности с представлением необходимых лингвистам знаков и букв.

Так продолжалось до 1983 г., пока дисплеи на жидких кристаллах не пришли на смену катодным трубкам. Сначала появился крошечный TANDY 100 с экраном на 4 строки текста, затем в 1984 г. появился Sharp PC 5000 с экраном на 8 строк, с графическими возможностями, 128К памяти и с процессором 8088, на котором работала операционная система MS-DOS, v. 2. Наступила эпоха серьезной разработки лингвистически ориентированного программного обеспечения.

Решение этой задачи потребовало значительных усилий и иногда даже вызывало разочарования. Не хватало ни времени на разработку, ни памяти, чтобы создать программы, которые было бы удобно использовать. В результате многие лингвисты испытывали большие трудности или избегали использовать компью-

теры. Так было до тех пор, пока компьютер не стал похож на значительно усовершенствованную пишущую машинку. Начиная с 1985 г. все большую роль среди персональных компьютеров стал играть Apple Macintosh. Но поскольку фирма Apple не давала другим производителям лицензии на свои разработки, это послужило препятствием для создания портативных компьютеров данной технологии и открыло путь к разработке портативных компьютеров, совместимых с персональными компьютерами фирмы IBM. Хотя во многих отношениях компьютеры Apple Macintosh превосходят компьютеры типа IBM PC, отсутствие портативных компьютеров типа Macintosh с автономным питанием привело к тому, что лингвистически ориентированное программное обеспечение начали разрабатывать для IBM-совместимых машин. Среди сотрудников SIL большинство имеет портативные компьютеры с экраном CGA. Это означает, что для изображения специальных знаков необходимо разрабатывать программы, работающие в графическом режиме. Сейчас появились улучшенные экраны, так что эта проблема понемногу снимается. Иногда лингвисты жаловались на неудобство программного обеспечения. И только в январе 1991 г. я стал получать благожелательные отзывы на программы SIL от лингвистов, испытывающих трудности в обращении с компьютером.

ЛИНГВИСТИЧЕСКИ ОРИЕНТИРОВАННЫЕ ПРОГРАММЫ SIL

Большинство программ, используемых сотрудниками SIL в полевых исследованиях, написано самими же сотрудниками SIL специально для этих целей. Иногда используются также коммерческие программы и программы других лингвистов. Сначала я рассмотрю коммерческие программы.

Microsoft Word — это коммерческий редактор, широко используемый в SIL. Поскольку сотрудники SIL в основном применяют экраны CGA здесь была разработана собственная технология получения специальных знаков в Word'e, а также технология модификации драйверов принтеров для Word'a. Моя основная трудность состоит в том, что файлы Word'a несовместимы с другими программами, которыми пользуются лингвисты.

Электронные таблицы — другой тип используемых лингвистами коммерческих программ. Такие программы применяются при сопоставлении, заполнении и обработке анкет для различных тестов, а также при накоплении данных по обучению грамотности (от учета книг до учета успеваемости).

Из коммерческих программ также широко известна программа для настольной полиграфии "Publish it!" Таймворкса. Эта программа применяется для издания учебников. В Папуа-Новая Гвинея SIL работает со 140 из 800 языков, и издание дешевых книг на этих языках представляет серьезную проблему. Сейчас небольшие книги часто готовятся прямо на компьютере, который осуществляет набор, верстку и иллюстрирование. Лингвист заменяет английский текст его переводом на местный язык и делает распечатку оригинал-макета книги.

Ventura Publisher — наиболее мощный программный продукт для этих целей.

ВСПОМОГАТЕЛЬНЫЕ ЛИНГВИСТИЧЕСКИЕ ПРОГРАММЫ, РАЗРАБОТАННЫЕ В SIL

Вспомогательные лингвистические программы (утилиты) — это программы для обработки любых текстовых файлов, не обязательно в лингвистических целях. Такие программы по своей природе имеют обобщенный характер и используются совместно со специфическими лингвистическими программными средствами.

Имеются два главных типа таких программ, а также множество отдельных программ специального назначения (например, для создания специальных служебных файлов или для преобразования различных файловых форматов).

Первый тип имеет название "пакет DTS" (Direct translator support). Он включает в себя многоязычный редактор ED, язык для таблиц подстановок CC (Consistent changes), форматтер текстов MS (manuscripter). Кроме того, сюда же входят программы DESIGN, SHAPES, FONTCO, KEYDEF и KEYSWAP, предназначенные для настройки клавиатуры и подготовки специальных знаков для экрана и принтеров.

Второй тип программ предназначен для анализа текстов в пакетном режиме. Эти программы образуют пакет TA (Text analysis).

ПАКЕТ DTS

Текстовый процессор (редактор) выполняет две главные функции: во-первых, с его помощью осуществляется экранное редактирование текста, а во-вторых, он печатает текст в нужном формате. В пакете DTS эти функции выполняются двумя отдельными программами — ED и MS.

ED может использоваться как простой текстовый редактор, но он обладает также мощными средствами работы со специальными знаками. Он может переключаться в графический режим для загрузки экранного шрифта и, конечно, использовать загружаемые шрифты при работе с экранами EGA и VGA. Кроме того, он может использовать и собственный метод обработки специальных знаков, при котором специальный знак представляется одним или несколькими стандартными. Сложные знаки могут образовываться путем наложения на буквы диакритик. Таким образом преодолевается ограничение компьютерного алфавита 128 или 256 стандартными знаками, и могут быть образованы слоговые графемы, например, для амхарского.

ED не является специализированным английским редактором, в нем отсутствуют средства для проверки орфографии и словарь синонимов, характерные для современных редакторов (текстовых процессоров). Существующая версия ED не может обрабатывать текст справа налево, но при использовании специального файла возможно преобразование текста на вводе. Новая версия ED позволяет определять знаки любого разумного размера, что используется в буквах со многими диакритиками, например, в деванагари в буквах с особо сложным начертанием.

Специальные знаки изготавливаются программами DESIGN и SHAPES, входящими в пакет DTS. Распознавание таких знаков, представляемых на вводе буквосочетаниями, осуществляется с помощью дополнительного файла ED.LRN.

ED может обрабатывать одновременно только один файл, его величина определяется доступной памятью.

DTS содержит альтернативную по отношению к файлу ED.LRN программу KEYSWAP, которая, будучи резидентной, отслеживает комбинации клавиш, изображающих сложные знаки. KEYSWAP настраивается с помощью программы KEYDEF. Она может обслуживать до девяти схем клавиатурного процесса.

CC может использоваться в простейшем случае для того, чтобы заменять знаки или их комбинации на специальные знаки. Но могут быть определены и более сложные контекстно зависимые подстановки. Например, последовательность пробелов заменяема на единственный пробел, причем есть возможность удалять пробелы и изменять в начале строки формат файла. Этот процесс управляется таблицей замен, которая может быть создана в виде обычного текстового файла с помощью текстового редактора, а затем откомпилирована.

MS может обрабатывать очень сложные специальные знаки, создаваемые программами DESIGN, SHAPES и FONTCO. Она печатает текст на принтере в графическом режиме или загружает в принтер нужный алфавит. MS управляется командами форматирования, располагаемыми в обрабатываемом файле. Сами эти команды могут быть определены в таблице подстановок в форме:

обратный слэш (\), за которым следуют одна или несколько (до четырех) букв и пробел. Удобство этого метода состоит в том, что такие команды могут быть определены динамически.

Поскольку в MS почти полностью реализована СС, возможна динамическая (во время печати) подмена формата. Например, файл, подготовленный латинскими или кириллическими буквами, может быть отпечатан как правильный арабский текст.

ПАКЕТ ТА

В пакете имеются пять программ: SRT, WL, DICR, IC и WS.

SRT — это очень мощная сортировка, WL предназначена для подготовки словоуказателей, DICR — обратных словарей, IC — конкордансов в интерактивном или в пакетном режиме. Программа WS сегментирует слова.

Программа SRT настраивается путем подготовки специального управляющего файла, в котором содержатся такие параметры обработки, как специальные знаки, буквосочетания, порядок сортировки слов и т.д. Для подготовки такого ans-файла используется программа ANSQ.

Пример:

Однажды пришлось сортировать немецкий словарь так, чтобы гласные с умлаутом были отсортированы, как будто они являются буквосочетаниями соответствующей буквы с буквой *e*. Это было сделано с помощью командного файла, в котором сначала выполнялась программа СС, затем — SRT, затем снова СС. Первое выполнение программы СС заменяло буквы с умлаутом на нужные буквосочетания, например, *й* на *%ие%*. Программа сортировки была настроена так, чтобы она игнорировала букву *%*. Наконец, при втором выполнении программы СС были произведены обратные подстановки.

Параметры для программ СС и SRT могут быть заданы в программной строке. Таким образом, весь процесс автоматизируется одним командным файлом.

СПЕЦИАЛИЗИРОВАННЫЕ ЛИНГВИСТИЧЕСКИЕ ПРОГРАММЫ

Ниже я рассмотрю семь главных программ и пакетов программ, разработанных в SIL специально для его сотрудников-лингвистов. Было бы неправильно понимать это так, будто эти программы были разработаны специально под идеологию SIL. Каждая из программ реализует представления определенного сотрудника, имеющего большой опыт полевых исследований, и поэтому решает комплекс специфических проблем. Это не значит, что авторами программ были лингвисты.

Первые пять программ работают на IBM-совместимых компьютерах и перечисляются в том порядке, в котором они, возможно, использовались бы в полевом исследовании. Это программы WORDSURV, CECIL (и SPECTRUM), FINDPHON, SHOEBOX и FIESTA.

Шестая программа, помощник переводчика, предназначена для машин Apple Macintosh и опирается на использование программы HYPER-CARD. Одним из стимулов использования машины Apple Macintosh было удобство обработки на ней арабских текстов.

Седьмая программа, скорее пакет программы, также предназначенных для IBM-совместимых машин, называется CADA (Computer aided dialect adoption). Она предназначена для преобразования текста на одном диалекте в текст на близкородственном диалекте.

WORDSURV

WORDSURV (текущая версия 2.4) написана сотрудником SIL Д. Уимбишом (John Winbish). Она предназначена для лексикостатистического и фоностати-

стического анализа сопоставительных списков слов сравниваемых языков или диалектов.

Если сравниваются десять языков или диалектов, то возможны 45 сопоставительных вариантов. Если необходим один час, чтобы провести лексико-статистический анализ любой пары языков, то необходима неделя, чтобы проанализировать все 45. Анкета, содержащая 50 списков, потребовала бы для обработки 1275 вариантов 7 месяцев. При использовании WORDSURV основное время уходит на ввод данных в компьютер, например, 10 списков вводятся за 10 часов, а 50 — за 50 часов. Таким образом, анкета, содержащая 50 списков, может быть обработана за семь дней вместо семи месяцев, причем компьютер ведет расчеты тогда, когда исследователь отдыхает.

К сожалению, WORDSURV не работает с CGA, поэтому специальные знаки можно обрабатывать только на EGA и VGA.

CECIL

CECIL (текущая версия 1.2A) и соответствующая программа SPECTRUM написаны Ф. Брассетом (Philip Brassett) специально для SIL. Программа работает со специальным оборудованием CECIL INTERFACE 55 (CI-500), именуемым в обиходе "CECIL-box". Этот прибор связывает магнитофон с компьютером, что создает походную фонетическую лабораторию с автономным питанием.

CECIL оцифровывает и запоминает звуковую речь. Скоростью обработки речевого сигнала управляет программа, обычно она равна 20 килогерцам, что позволяет записывать трехсекундные кванты речи. Лингвист имеет возможность наблюдать осциллограмму речевого сигнала и просмотреть любой ее участок вплоть до одного цикла. Длительность звука измеряется с точностью до одной тысячной секунды. Речевой сигнал или его часть воспроизводятся с нормальной скоростью, с замедлением в три и в пять раз. Кроме того, выдаются графики изменения громкости, частоты колебания голосовых связок, изменения качества звука с цифровой индикацией частоты и громкости.

Графики, выдаваемые CECIL, могут совмещаться. Таким образом, например, изменения громкости и частоты сравнимы, и наблюдаемы зависимости, если они существуют. Два высказывания могут быть также наложены друг на друга. Так сравнивают речевые сигналы различных дикторов или одного и того же диктора.

CECIL может использоваться этномузыковедами, хотя и не была сконструирована для этих целей. В этом случае на эксперимент накладывается ряд ограничений, так что исследоваться может только сигнал солиста без аккомпанемента либо сигнал одиночного инструмента без аккордов в полосе от 50 до 500 герц длительностью до 10 сек. (в этномузыковедении нужен анализ звучания длительностью, скажем, одну минуту).

SPECTRUM получает на вход бинарные файлы речевых сигналов, выдаваемые CECIL, вычисляет спектры частот всего сигнала или его части. При расчете спектрограмм программа совершает до 75 000 вычислений на сигнал длительностью в одну секунду (в не более чем 250 интервалах по 300 вычислений в каждом). Неудивительно, что такое количество вычислений может длиться минуты в зависимости от быстродействия компьютера и настройки программы. И CECIL, и SPECTRUM удобны в использовании, имеют удобные меню, обеспечивают как общее, так и ситуативно-контекстные объяснения и подсказки (help). Язык меню может быть изменен без recompilation программ. Имеются три учебника, один из которых называется "Сбор и акустическая интерпретация фонетических данных".

Обе программы работают на компьютерах с экранами CGA, DSCGA и EGA. "CECIL-box" имеет автономное питание на батареях, подключаемых к параллельному порту, имеет стандартный, поставляемый вместе с ним интерфейс, выдает байтовую оцифровку и может обрабатывать не только сигналы,

записанные в полевых условиях, но и студийные записи. Для использования необходим матричный принтер, совместимый с принтерами типа Epson или Toshiba, предпочтительно 24-игольчатый.

Этот прибор изготавливается в SIL, так как не существует приемлемых промышленных продуктов такого назначения; если таковые появятся на рынке, в SIL, видимо, напишут для них новые программы, аналогичные описанным выше.

FINDPHON

FINDPHON (FF, текущая версия 5) написана сотрудником SIL Д. Бивеном (David Bevan), и хотя она до сих пор не совсем готова, она способна выполнять некоторые весьма полезные функции, встречающиеся в фонологическом анализе. Полная разработка программы будет, видимо, завершена в ближайшие месяцы.

Программа принимает на вход транскрибированные фразы или отрывки текстов вместе с различного рода примечаниями к ним, такими, как код записи, ударения, перевод, классификационный идентификатор и т.п.

Программа проста в эксплуатации, имеет развитую систему меню, встроенный редактор, работающий со знаками Международной фонетической транскрипции. Редактор обладает мощными средствами контекстного поиска, приспособленными для целей фонологического анализа; он, например, имеет возможность поиска контекста с точностью до класса (скажем, по аргументу F для поиска фрикативных). Результаты могут быть отредактированы, отсортированы и напечатаны.

Сейчас разрабатываются процедуры поиска минимальных пар и генерации контрастивных оппозиций.

FF работает на машинах с экранами CGA, DSCGA и VGA и требует использования 24-игольчатого принтера, совместимого с принтерами типа Epson или Toshiba.

SHOEBOX

SHOEBOX (текущая версия 1.2A) написана сотрудником SIL Дж. Уимбишом (John Wimbish). Сначала она была задумана как программа для работы с лексическими базами данных, но позднее была дополнена такими функциями, как ввод подстрочника и ведение картотек. Программа ведет до семи баз данных одновременно и рабочий текстовый файл. При вводе подстрочника образуются один или несколько словарей и выбираются данные из нескольких баз. Одновременно в другие базы данных заносятся данные, например, грамматические пометы или примечания.

Сейчас базы данных индексируются только по одному ключевому полю, но этот недостаток компенсируется двумя достоинствами: мощным групповым поиском и переменной длиной записи.

SHOEBOX используется также при анализе антропологических или социолингвистических данных.

Программа имеет весьма дружелюбный пользовательский интерфейс с развитой системой меню и help'ами, снабжена хорошим учебником, где описывается не только сама программа, но и возможные лингвистические процедуры над накоплением и использованием этой программы материалом.

SHOEBOX работает на машинах с экранами CGA, EGA или VGA. С экраном CGA возможно переключение в графический режим, таким образом, на всех экранах воспроизводятся специальные знаки. Поскольку SHOEBOX использует большое количество данных, она требует много оперативной и дисковой памяти, и только для очень небольших применений могут использоваться гибкие диски.

FIESTA (текущая версия 4.5n) написана сотрудником SIL Дж. Олсопом (John Alsop). Эта программа предназначена для быстрого поиска, просмотра, сравнения и редактирования больших текстов, которые могут быть разбиты на субтексты переменной длины (не более 2500 байтов). Такие субтексты, например, предложения, рассматриваются как отдельные записи и снабжаются специальными маркерами. Каждая запись содержит до 180 помеченных полей.

FIESTA — важный инструмент лингвистического анализа, поскольку обладает интерактивностью, быстрым откликом. На базе данных объемом 1 Мб. поиск длится доли секунды на достаточно быстром компьютере. Кроме поиска на больших текстах FIESTA используется также для индексирования словарей и библиографических картотек. Программа может индексировать каждое слово документа и запоминать его в двух байтах. Программа работает с файлами до 180 Кб. Большие файлы индексируются кусками по 180 Кб. и затем сливаются. Это довольно медленная процедура, но в будущем мы надеемся ее ускорить.

Файлы с подстрочниками, создаваемые программой SHOEBOX, могут быть проиндексированы программой FIESTA и благодаря этому обеспечивается более быстрый доступ к файлам.

FIESTA первоначально предназначалась для лингвистов, занимающихся переводом Библии. Им были нужны мощные средства проверки перевода, но программа предусматривала и другие процедуры: 1) поиск и представление на экране контекстов заданных слов, фраз или цепочек знаков, включая контексты разрывных слов; 2) представление на экране полного или частичного списка слов к тексту или списка слов, содержащих заданные буквосочетания; 3) представление списка слов с заданной слоговой структурой; 4) подсчет частот заданных слов, фраз или цепочек знаков; 5) контекстнообусловленное изменение написания или удаление из всего текста заданного слова или части слова, замена целых фраз; 6) сортировка записей по заданному полю, переписывание знаков в полях в обратном порядке, что позволяет получать обратные словари.

FIESTA работает на машинах с экранами CGA, EGA или VGA, причем в CGA возможно переключение в графический режим, что обеспечивает обработку специальных знаков.

FIESTA работает одновременно только с одной базой данных, причем каждая должна находиться в своей субдиректории, но поскольку базы данных могут быть велики, много больше 1 Мб., и данные должны храниться на диске, нужен жесткий диск.

ПОМОШНИК ПЕРЕВОДЧИКА

Помощник переводчика был разработан для переводчиков, которым нужен арабский шрифт вместе с английским и другим алфавитом. Он позволяет переводчикам работать на экране с параллельными текстами, а также накапливать словарные данные, например, содержащие решения об определенных написаниях или переводных эквивалентах.

CADA

CADA (computer assisted dialect adaptation) — пакет программ, разработанных сотрудниками SIL, преимущественно докторами Б. Манном (Bill Mann) и Д. Вибером (David Weber) с целью упрощения адаптации текстов на одном языке или диалекте к восприятию носителями другого. Первоначальные эксперименты были проведены среди носителей многочисленных диалектов высокогорных индейцев кечуа в Перу. Записи на различных диалектах были использованы для реконструкции протокечуа, и адаптированные тексты написаны на этом протоязыке. Поскольку кечуа — агглютинативный язык,

в программе имеется процедура идентификации морфем и их корректной подстановки формами других диалектов. В дальнейшем планируется создать процедуру синтаксических преобразований, необходимую для многих неагглютинативных языков.

Пакет состоит из двух главных программ, AMPLE и STAMP. Первая программа анализирует входной текст на исходном диалекте, а вторая выдает синтез текста на целевом варианте. Для каждой из программ имеется свой учебник; есть также обучающий материал для совместного использования обеих программ.

НЕОБХОДИМЫЕ ПЕРСПЕКТИВНЫЕ РАЗРАБОТКИ

Лингвисты SIL получили определенные преимущества от использования большого количества мощных программных средств, но это вовсе не значит, что такие средства имеются для всех возможных областей применения ЭВМ. Имеются три области, для которых явно не хватает программных средств.

Одной из таких областей для разработок является специализированный программируемый редактор для записи антропологических и других данных. Было бы хорошо, если бы такой редактор работал совместно с программой, подобной SHOEBOX, чтобы в нем были реализованы и другие возможности, но при этом в SHOEBOX нужно было бы улучшить алгоритм группового поиска данных.

Программируемый спецредактор должен иметь возможность задавать на экране вопросы пользователю и предоставлять в меню варианты ответов, а также принимать решения о следующих далее вопросах в зависимости от полученного ответа.

Предварительная версия такого редактора уже существует и однажды была использована в Восточной Африке для записи словаря самими информантами. Информант получал инструкцию: "Введите слово". После этого ему задавались все новые и новые вопросы, отвечая на которые информант заставлял редактор формировать очередную словарную статью в нужном компьютерном формате, о котором пользователю ничего не было известно и о котором он мог и не думать.

Второй областью применений является программа для отображения схем дискурса. В SIL имеются две такие программы, а именно, CHARTER и ALLIGN, но я не уверен, достаточно ли они удобны и развиты.

Третьей областью является конструирование компьютерных букварей, и я надеюсь что такая программа скоро появится.

ОБУЧЕНИЕ ПОЛЬЗОВАТЕЛЕЙ

Хотя программы и должны быть ориентированы на потенциальных пользователей, не следует думать, что все или хотя бы большинство лингвистов могли мыслить как программисты. Поэтому для лингвистов нужно разработать методику первичного обучения использованию компьютеров.

Многие лингвисты, включая работников университетов, до недавнего времени избегали пользоваться компьютерами. Поэтому первой задачей в обучении лингвистов является завоевание доверия и передача им начальных знаний компьютерного дела. Вторая задача — обучение пользованию одной или двумя нужными им программами, чтобы постепенно сложилось представление о практической пользе компьютеров.

Лингвисты заняты вопросами лингвистического анализа фактов языка, а не вопросами функционирования программного обеспечения. В большинстве же учебников обсуждаются вопросы, как использовать соответствующие программы, а не задачи лингвистического анализа. К счастью, это положение начинает меняться.

Сейчас, когда программы приобрели дружественный интерфейс, а компьютеры подешевели, все большее число лингвистов используют их для решения профессиональных задач, а не только для редактирования текстов.

НОВЫЕ РАЗРАБОТКИ КОМПЬЮТЕРОВ И ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ

В последние годы рынок компьютеров был и интересен, и несколько хаотичен, такое положение сохранится и далее. Запад переживает период экономического спада. Люди не покупают новых компьютеров. С другой стороны, на рынке появились новые типы компьютеров. Более того, компьютеров на рынке стало больше, чем потенциальных покупателей. Поэтому цены быстро падают и будут падать. Даже такие большие компании, как IBM и Compaq, сократили цены на 25%.

Для тех, кому нужен портативный компьютер с автономным питанием, есть новая модель 386SX размером с блокнот (30 × 26 × 5 см), с жестким диском на 20 Мб. с VGA-экраном на жидких кристаллах, с 1 Мб. оперативной памяти, с батареей на 2—3 часа работы, весом около 3 кг и стоимостью от 2500 до 3000 долл. Эквивалентный настольный компьютер можно купить за две трети этой цены. Эти компьютеры уже не могут стать меньше, они снабжаются стандартной клавиатурой и экраном подходящего для работы размера. Но они будут продаваться с еще большей оперативной и дисковой памятью, с процессорами, требующими еще меньше энергии.

Старые портативные компьютеры с экраном CGA на жидких кристаллах, с одним или двумя гибкими дисками продаются по цене менее 700 долл.

Есть и новые перспективные разработки. Для IBM совместимых компьютеров на рынок поступает графический пользовательский интерфейс Windows 3.0, а скоро и Windows 3.1, операционная система MS DOS 5 и новые маломощные (экономные) процессоры. Фирма Apple наконец-то приступила к производству дешевой версии своего компьютера Macintosh, и теперь в половине ее продаж будет этот компьютер.

CELLAR — НОВАЯ ГЕНЕРАЛЬНАЯ РАЗРАБОТКА

Сотрудник SIL Г. Саймонс (Gary Simons) возглавляет разработку нового программного обеспечения для полевых исследований. Новый проект CELLAR (Computing environment for linguistic, literacy and anthropological research) представляет собой попытку связать воедино все ранее разработанные программные процедуры и методы, чтобы увеличить их мощность и придать им большую согласованность. Срок разработки первой полной версии — январь 1995 г.

Ядром разработки является язык программирования высокого уровня, специально предназначенный для лингвистических задач. На этом языке можно будет программировать ввод и отображение большого набора алфавитов, накопление в базах данных лингвистических данных различных типов, включая гипертекстовые связи, дружественный пользовательский интерфейс, обработку фонетических данных.

Это ядро представляет собой оболочку, под которой можно будет писать конкретные прикладные программы, при этом для лингвистов должно обеспечиваться удобство, понимание вычислительной ситуации и быстрый доступ к данным.

CELLAR можно будет использовать только на более мощных компьютерах, чем те, которыми мы обычно пользуемся, — с более быстрыми процессорами, с большей памятью, с экранами с большей разрешающей способностью и с более емкими жесткими дисками. Портативные компьютеры, о которых мы говорили выше, постепенно развиваются в этом направлении, но все-таки это не те компьютеры, на которых будет работать CELLAR.

SIL счастлив сделать свое программное обеспечение доступным академическим учреждениям по вполне умеренной цене. Во многих случаях мы счастливы даже подарить его, но не хотелось бы, чтобы это было понято как обязательство.

Поскольку условия, в которых работают сотрудники SIL, весьма различны, наше программное обеспечение также довольно гибко, приспособляемо к различным задачам и алфавитам, но в то же время оно должно быть простым и удобным в использовании. Эти требования, конечно, конфликтуют друг с другом, но нам в большинстве случаев удавалось достигнуть разумного компромисса.

Было бы интересно узнать, какое программное обеспечение появится в ближайшие годы, когда пакет CELLAR будет готов к широкому использованию.

Иногда я мечтаю о программном обеспечении и технике, которые были бы в моих руках, если бы те одиннадцать лет, которые я провел в полевых исследованиях в Гане, случились лет на пятнадцать позже. Тогда бы многие вещи я сделал лучше. Программы SIL уже внесли свой заметный вклад в лингвистические исследования ученых многих стран, и этот вклад будет несомненно расти.

АДРЕСА SIL

Euroasia Area Director, SIL, Horsleys Green, High Wycombe, HP14 3XL, England.

SIL Executive Vice-president, International Linguistics Center, 7500 W. Camp Wisdom Road, Dallas, TX 75236, USA.

Перевел с английского *Андрющенко В.М.*