Локализация для Китая и Японии

Фрэнк Лин (Frank Lin) и Ангелика Церфасс (Angelika Zerfaß)

Экономики Китая и Японии занимают второе и третье места в мире по величине и обладают огромным рыночным потенциалом для иностранных продуктов. В Китае больше всего пользователей компьютеров и Интернета. Несмотря на свои достижения в экономике, экспорте и технологиях Япония импортирует довольно много технологий разработки программного обеспечения.

Однако многообещающие рыночные возможности омрачаются сложностями интернационализации программного обеспечения для Японии и стран, где говорят на китайском языке. Интернационализация для этого региона считается более затруднительной, чем для Европы, преимущественно из-за технической необходимости многобайтовой кодировки языков.

Наряду с кодировкой существуют и менее очевидные, хотя не менее распространенные и сложные проблемы, связанные с различиями языковых и культурных норм. Проекты локализации для этих двух языков обычно более затратные, требуют больше времени для реализации и сопряжены с большими объемами технических работ и тестирования.

Лингвистические проблемы

Пожалуй, самым сложным аспектом интернационализации программного обеспечения для японского и китайского языков является их письменность. В китайском и японском языках существуют похожие иероглифические системы: ханьцзы и кандзи соответственно. Кроме иероглифов, в японском языке используется две системы алфавитного письма, хирагана и катакана, которые вместе называются кана. Основная причина, по которой в стандартах кодировки пришлось ввести поддержку многобайтовых символов, заключается в том, что китайских иероглифов слишком много. На данный момент в Юникоде закодировано свыше 40 000 китайских иероглифов. Большинство этих иероглифов не используется в повседневной жизни, тем не менее, их поддержка программным обеспечением крайне важна, поскольку обычно эти редкие иероглифы используются в именах людей, а также названиях объектов и мест.

Невзирая на схожесть письменности этих двух языков, лингвистически они очень отличаются. Китайская грамматика относительно проста, и в ней обычно не наблюдается морфологических вариаций. Глаголы не имеют категорий вре-

мени, числа и лица. Существительные не имеют формы множественного числа и в большинстве случаев — категории рода. Прилагательные не имеют флексий. Японская грамматика, наоборот, более сложная. Японские глаголы, как и в европейских языках, имеют категории времени, числа и лица. И хотя у существительных нет формы множественного числа, как и в китайском языке, прилагательные изменяются по родам, числам и падежам. С точки зрения структуры предложений в китайском используется порядок слов «подлежащее — сказуемое — дополнение», как и в английском языке, в то время как в японском применяется модель «подлежащее — дополнение — сказуемое».

И в китайском, и в японском языке традиционно пишут как по вертикали, так и по горизонтали. При письме по вертикали иероглифы пишутся сверху вниз, а следующая строка записывается слева от текущей (справа налево). Горизонтальная ориентация стала доминировать после Второй мировой войны. В этом случае иероглифы пишутся слева направо или справа налево, но на сегодняшний день чаще используется вариант «слева направо», как в английском языке, что упрощает адаптацию китайского и японского языков к современным компьютерным системам.

На китайском языке сложно научиться писать из-за большого количества иероглифов и их сложной структуры. Чтобы упростить изучение, в 50–60-х годах двадцатого века в Китае было проведено несколько орфографических реформ, в результате которых упростилось написание многих общеупотребительных иероглифов. Таким образом было создано упрощенное китайское письмо. Традиционное китайское письмо используется на Тайване, в Гонконге и Макао. В Сингапуре сейчас принято упрощенное письмо, хотя иногда для написания имен людей здесь используется традиционное китайское письмо. В упрощенном и традиционном китайском письме просто используются разные наборы символов, а не другой язык. Тем не менее, в некоторых ситуациях наблюдаются региональные отличия в лексике, особенно это касается компьютерной терминологии. Возьмем, к примеру, слово *компьютер*. В Китае это 计算机. То же слово при использовании традиционного китайского письма пишется как 計算機, что фактически означает *вычислительное устройство*.

По мере глобализации культуры и торговли, а также с распространением Интернета в странах, где говорят на китайском языке, стала возрастать потребность в поддержке обоих видов письменности в одном и том же программном приложении или на одном веб-сайте. Хорошим примером могут служить страницы Википедии на китайском языке (**рис. 1**), где в раскрывающемся меню можно выбрать упрощенное или традиционное письмо. При этом для традиционного письма предлагается два диалекта: для Тайваня и для Гонконга/Макао, а для упрощенного письма — версия для Китая и Сингапура.

В европейских языках используется алфавитная письменность, а для сортировки слов применяется строго заданный алфавитный порядок. При этом иногда вспомогательную роль играет написание

Рис. 1. Поддержка упрощенного и традиционного письма на китайском языке в Википедии

с прописной или строчной буквы, а также диакритические знаки. Обычно используется лексический порядок сортировки, хотя в некоторых случаях бывают небольшие вариации, например использование международной сортировки для испанского языка и сортировки телефонной книги для немецкого.

Сортировка для японского и китайского языков неочевидна, поскольку в этих языках отсутствует чисто алфавитная письменность. За много лет до появления компьютеров специалисты пытались найти и стандартизировать эффективный способ сортировки. Одним из самых популярных методов является сортировка по корню слова, по числу штрихов и по произношению. В материковом Китае ее называют *пиньинь*, а на Тайване — *бопомофо* или *чжуинь*. Сортировку по произношению также называют фонетической. В Китае преобладает сортировка пиньинь, хотя определение сортировки — процесс многоуровневый. Например, основным уровнем может быть пиньинь. Если происходит совпадение (вследствие одинакового произношения многих китайских иероглифов), можно использовать второй уровень, например число штрихов в иероглифе, и так далее. Для записи пиньинь используется латинский алфавит с диакритическими значками и четко определенной сортировкой.

Сортировка в японском языке еще сложнее, чем в китайском. Японские слова и предложения пишутся с использованием трех письменных систем плюс латинские символы, которые называются *ромадзи*. Хотя в пределах хираганы и катаканы правила сортировки заданы четко, сортировка кандзи однозначно не определена. На **рис. 2** показано японское предложение с использованием всех систем письма.

Как и в китайском языке, в японском наиболее распространена фонетическая сортировка иероглифов. Произношение иероглифов кандзи может обозначаться буквами кана. Такое обозначение называется *йоми* или иногда *фуригана*. Например, написанное кандзи слово 鈴木 можно представить с помощью букв кана すずき (судзуки). Четкой фонетической сортировки можно добиться, если все кандзи в словах или предложениях преобразовать в кану (эквивалентные буквы в хирагане и катакане можно относить к одному рангу), то есть, по сути, нормализовать иероглифы.

Проблема заключается в том, что произношение символов кандзи в японском языке не фиксировано. Слово 鈴木 произносится как すずき (судзуки),

Рис. 2. Японское предложение с использованием всех письменных систем

私	の	名前	は	アンゲリカ	(anger ika)	です
Ватаси	но	намаэ	ва	Андзерика		десу
Я	Частица, используемая для преобразования местоимения *я* в местоимение *мое*	имя	Иероглиф *ха* произносится как *ва* для обозначения предмета разговора	Анжелика (в японской фонетической системе звуки *р* и *л* передаются звуком *р*)	Транскрипция имени на *ромадзи*	есть

но точно так же произносится и слово 錫気. С другой стороны, символы поразному произносятся в зависимости от контекста, в котором они используются, что тоже не в пользу сортировки. В слове 鈴木 второй иероглиф кандзи 木 произносится как き (ки). В слове 木本 иероглиф 木 произносится как もく (моку). Причина взаимно неоднозначного соответствия между написанием кандзи и произношением кроется в заимствовании китайской письменности в японские слова с сохранением японского произношения.

Как станет ясно из приведенного ниже обсуждения технических деталей, «проблема произношения кандзи» оказывает большое влияние на все технические работы, которые необходимо выполнить для правильной сортировки в японском языке. Фонетическая сортировка для японского языка не всегда внедряется в программные приложения из-за сложности записи правильного представления кандзи в виде кана и необходимости менять алгоритмы сравнения иероглифов.

Как отмечалось ранее, для сортировки можно использовать фуригану. Иногда ее также применяют для обозначения произношения иероглифов, которое указывается над самими иероглифами. В этом случае для них используется шрифт «Агат». Ниже приведены примеры написанных шрифтом «Агат» иероглифов на китайском и японском языках.

běi jīng	とうきょう
北京 Слово *Пекин* на китайском языке	東京 Слово *Токио* на японском языке

Культурные проблемы

Культурный аспект интернационализации программного обеспечения связан с двумя вопросами: культурными особенностями представления и использования программного обеспечения и местными нормами. В программном обеспечении должна быть предусмотрена определенная гибкость и способность к адаптации.

Примером культурной локализации может служить использование символов. Знак красного креста часто применяется для обозначения скорой помощи или больницы, в исламских странах для этого служит красный полумесяц. Использование знаков может носить как культурный, так и политический или географический характер. Для обозначения национального языка на пикто-

Рис. 3. Представление данных о погоде в трех странах

грамме или кнопке обычно используется национальный флаг, хотя флаг Англии редко используется в США для обозначения английского языка. Так же редко можно встретить флаг Китая, обозначающего китайский язык, на тайваньских веб-сайтах, за исключением случаев, когда он используется для различения упрощенного и традиционного китайского письма.

Кроме того, представление более сложной информации, такой как данные о погоде, может сильно отличаться в зависимости от ожиданий пользователей. Примеры на **рис. 3** взяты из прогнозов погоды на портале Yahoo! В Германии простые изображения облаков, капель дождя или солнца для каждого дня позволяют получить общее представление о местной погоде. В Соединенных Штатах Америки отображается карта страны со статическими символами в виде облаков и солнца. При этом для Японии также используется карта страны, но с анимированными изображениями.

Что касается программного обеспечения, восприятие изображений может повлиять на используемые для них значки. Одним из примеров служит справочная система. В Соединенных Штатах Америки или Европе используется изображение человека, похожего на Эйнштейна, волшебника или даже животного, например кота или дельфина, следуя указаниям которого, он может найти ответы на свои вопросы. Для азиатских стран, в которых животные находятся на более низкой ступени, чем человек, животное, которое учит человека, будет не самым лучшим выбором для значка справки. То же касается изображения волшебника в западном стиле, которое следует заменить азиатским персонажем, например монахом (хотя в этом случае нужно с осторожностью подойти к религиозной символике) или мудрым старцем.

Знаки также могут вызывать различные ассоциации. В западном полушарии значок ☑ может означать выполнение какого-либо действия или правильность утверждения, однако в Японии этот символ используется для обозначения ошибок в домашних работах и, со-

ответственно, имеет более негативный оттенок. Кроме того, если пользователь из Соединенных Штатов Америки или Европы допускает ошибку, программное обеспечение может предупредить его с помощью звукового сигнала. А вот в Японии, где много людей часто работает в небольших помещениях, звуки, сообщающие об ошибке, будут восприниматься не очень хорошо.

Еще одним ярким примером культурных различий восприятия могут служить цвета. Например, в китайской версии сайта Yahoo! С индексами финансового рынка красный цвет используется для обозначения цен на растущие акции (красный = праздничный), в то время как в версии для США красный применяется в прямо противоположном значении (красный = сигнал тревоги).

Кроме того, в азиатской нумерологии может использоваться другая символика и, следовательно, это необходимо учитывать в программном обеспечении. Некоторые числа символизируют успех (например, восемь в Китае), в то время как другие могут оказаться запретными (такие, как четыре во многих странах Азии, поскольку в китайском и японском языках число четыре по произношению напоминает слово *смерть*). Это ничем не отличается от попытки умышленно избегать числа 13, например при нумерации этажей в некоторых странах Запада.

В большинстве проектов локализации для европейских языков наиболее часто в программном обеспечении необходимо адаптировать к местным особенностям форматы чисел и дат. В этих проектах форматы даты связаны с порядком расположения месяца и числа, а числовые форматы — с десятичной запятой. В китайском и японском программном обеспечении все немного сложнее.

Самое большое отличие между западной системой счисления и системой счисления в Китае и Японии состоит в группировании цифр. В Соединенных Штатах Америки цифры объединяются в группы по три, в то время как в китайском и японском языках они объединяются в группы по четыре. В Китае и Японии для представления числа могут использовать либо арабские цифры, либо собственные числовые иероглифы.

Формат даты, характерный для этих стран, обладает любопытными отличиями, которые носят преимущественно геополитический характер. В Японии используется старинный восточноазиатский принцип календарного летоисчисления по «эпохам». Эпоха меняется, когда на престол восходит новый император, а год его восхождения является первым годом новой эпохи. Когда в стране начинается новая эпоха, отсчет лет начинается заново. Например, 2011 год в Японии будет записан как 平成23年 (*Хисэ, год 23*, где Хисэ — это царственное имя императора и название эпохи), а 1987 — как 昭和63年 (*Сева, год 63*). Одно из очевидных последствий использования другого календарного обозначения заключается в том, что в момент окончания эпохи требуется обновление операционной системы.

Среди стран, в которых говорят на китайском языке, Тайвань — единственный регион, где в календаре все еще используются названия эпох. Например, 2011 год на Тайване также называется

«100-й год Республики» (民國100 年). В отличие от Японии, название эпохи на Тайване не меняется. В Китайской Народной Республике к системе «нашей эры» перешли еще при ее основании в 1949 году.

В Японии и на Тайване наблюдается повсеместная адаптация календарной системы «нашей эры» в повседневной жизни и программных приложениях, особенно в Интернете. В этих случая даты записываются примерно вот так: 二〇一一年一月十 一日 (2011 + иероглиф, обозначающий год, + 1 + иероглиф, обозначающий месяц, + 11 + иероглиф, обозначающий день, = 11 января 2011 г.). Однако традиционные календарные системы по-прежнему необходимы, ведь для обозначения важных дат, например года рождения, люди обычно используют местный формат даты.

В именах людей в Китае и Японии, в отличие от западных языков, фамилия пишется перед именем: имя известного бейсболиста в действительности звучит как Судзуки Итиро, а не Итиро Судзуки. С технической точки зрения это может создавать проблему в тех программных приложениях, где используется западный порядок имен. На Западе только в Венгрии используется восточный порядок имен.

В ориентированных на США программных приложениях используются характерные для этой страны форматы представления данных: формат номера социального страхования, номера телефона и адреса. Стоит отметить, что в формате адреса обычно используется порядок от малого к большому: 1600 Pennsylvania Avenue, Washington, District of Columbia, USA (США). В Китае и Японии все наоборот. Типичным адресом будет такой: Китай, провинция Чжэцзян, г. Ханчжоу, ул. Чангань, номер 100. В типичном адресе в Японии может даже не указываться название улицы, а только город, конкретная часть города, ее район и номер дома: Япония, 107 Токио-то, Тиода-ку, Касумигасеки 1-3-2, г-ну Икута Масахару.

Проблемы, связанные с обозначением валют, возникают не только при локализации ПО для китайского и японского языков. Дело в том, что для обозначения самых важных валют в мире используется одиночный символ ($, €, £, ¥), который применяется как в Китае, так и в Японии. Тем не менее, для многих валют используются обозначения, состоящие из нескольких символов, в том числе и для тайваньской (NT$). Еще одна проблема, связанная с валютой, заключается в наличии дробных частей денежной суммы. В Японии и на Тайване отсутствуют деньги, мельче базовой валютной единицы. Программные приложения, в которых используются наименования валют, должны быть достаточно гибкими для решения этих двух проблем.

Пунктуация в китайском и японском языках аналогична пунктуации в английском. Но есть несколько существенных отличий: восточноазиатские квадратные кавычки (「 и 」), точка (。) и разделитель для иностранных имен. Барак Обама пишется バラク・オバマ на японском и 贝拉克 • 奥巴马 — на китайском, где точка в центре разделяет имя и фамилию в неазиатских языках. В обоих языках отсутствуют строчные и прописные буквы. В китайском нет ал-

фавита, а в японском алфавите не различают прописные и строчные буквы.

В целом, в японском и китайском языках отсутствует разрыв между словами в виде пробела. Предложение «Сегодня я планирую написать три рассказа», написанное на китайском языке с использованием упрощенного письма, будет выглядеть следующим образом: 今天我计 划写三本小说。А на японском так: 今日私は3つの小説を書くことを計画する。

Технические трудности

Поскольку для каждого восточноазиатского языка требуется поддержка многобайтовых символов для систем письма, обычно выбирается кодировка Юникод. Ряд наиболее распространенных средств разработки программного обеспечения и языки программирования совместимы с Юникодом. Тем не менее, множество программ (обычно это программы, созданные несколько лет назад) не поддерживает этот стандарт. Таким образом, при интернационализации программного обеспечения для Восточной Азии первоочередной задачей для этих программ, пожалуй, является обеспечение поддержки Юникода.

Однако в некоторых ситуациях внедрение поддержки Юникода — далеко не идеальное решение. Этому препятствуют технические и экономические причины. Программы без поддержки Юникода, в которых выполняется множество операций по обработке данных на уровне байтов и символов, при внедрении таковой становятся уязвимыми, зачастую из-за интерпретации данных. Примером может служить объявление переменной C++: char str[100]. Является ли строка иероглифов в программе строкой символов или строкой байтов? Строка байтов не должна превратиться в программе с поддержкой Юникод в строку двухбайтных символов. Эта вроде бы простая проблема обостряется с увеличением размера и возраста программы.

Для «устаревшего программного обеспечения» существует альтернатива, позволяющая отказаться от внедрения поддержки Юникода, хотя для материкового Китая это может быть проблематично из-за требования GB18030 (обязательный стандарт кодировки, установленный правительством Китая). Современную программу вполне возможно локализовать для китайского и японского языков без внедрения поддержки Юникода. В корпорации Майкрософт для отличных от Юникода кодировок существует так называемая многобайтовая кодировка MBCS. При использовании кодовой страницы для китайского (CP936 в Китае или CP950 на Тайване) или японского (CP932) языка для символа может применяться один или два байта. Символы в 7-битной таблице кодов ASCII (английские буквы и пунктуация) кодируются с использованием одного байта, а для всех остальных символов используется два байта (китайских символов, катаканы, хираганы и латинских букв с диакритическими знаками). В среде MBCS программирование осложняется тем фактом, что строка иероглифов теперь становится сочетанием однобайтовых и двухбайтовых символов, поэтому необходимо разработать функцию, которая позволит определять количество байтов

в следующем символе строки при ее синтаксическом анализе.

Поскольку в этих языках нет пробела, отделяющего слова или символы, программы, в которых предполагается, что разделителем слова является пробел, будут работать неправильно. Распространены также другие ошибочные допущения: использование строчных и прописных букв, западный порядок слов в именах и пунктуация.

В среде Microsoft Windows для восточноазиатских языков широко используется редактор методов ввода (IME). Пользователи могут быстро вводить китайские и японские иероглифы, а также кану, используя редактор IME и стандартную английскую клавиатуру. Пользователи компьютеров в материковом Китае используют в качестве устройства ввода английскую клавиатуру, поскольку слова на пиньинь записываются с помощью латинских букв. Пользователи на Тайване чаще всего применяют клавиатуру для традиционного письма китайского языка, на которой поверх клавиш с латинскими буквами нанесены фонетические знаки и корни слов. В Японии символы вводятся с клавиатуры либо в режиме каны, либо в режиме ромадзи. В режиме каны при нажатии клавиш выполняется ввод японской каны. В режиме ромадзи японские слова записывают с помощью латинских букв, а затем преобразовываются в кану.

Из-за необходимости сохранения обратной совместимости с устаревшими компьютерными системами японское программное обеспечение обычно должно поддерживать шесть типов кодировок. Наряду с тремя письменными системами используются три следующие схемы кодирования: полуширинная хирагана, полуширинная катакана и полуширинная буквенно-цифровая. Программное обеспечение должно распознавать эквиваленты в различных кодировках. Например, в Юникоде букве полноширинной катаканы ｱ соответствует кодовая точка 0x30A2. Кодовая точка полуширинной буквы ｱ — 0xFF71. Их обе необходимо стандартизовать.

Недетерминированный характер произношения кандзи в японском языке усложняет поддержку сортировки в компьютерах. Обычные методы сортировки, такие как таблица соответствия, неприменимы. Вместо этого часто необходимо хранить отдельно произношение каждого элемента данных кандзи, подлежащих сортировке. Ключом сортировки для соответствующих данных кандзи становится йоми. Йоми для кандзи можно зафиксировать явно и неявно. При явной фиксации пользователю просто предлагается ввести йоми для соответствующего иероглифа кандзи. При неявной фиксации программное обеспечение должно взаимодействовать с редактором IME для получения данных о нажимаемых клавишах в процессе ввода иероглифов кандзи с помощью IME.

Из-за сложности и масштаба работ при повторном проектировании фонетической сортировки и данных йоми в программном обеспечении на английском языке сортировка внедряется в программное обеспечение не всегда правильно. Примером великолепной реализации сортировки для японского языка может служить приложение Microsoft Excel.

Хотя для сортировки в китайском языке также используется произношение, оно, в отличие от японского языка, четко определено для большинства китайских иероглифов. Иными словами, у каждого иероглифа есть фиксированный ранг в системе сортировки пиньинь. Следовательно, для сортировки можно использовать таблицу соответствия. Современные системы баз данных обычно обеспечивают для китайского языка поддержку сортировки по принципу сопоставления. Существует небольшое количество китайских иероглифов, которые могут иметь несколько вариантов произношения. В таких случаях может потребоваться определенная настройка. В худшем случае можно рассмотреть внедрение стиля йоми, хотя для большинства приложений это лишает перевод на китайский язык всякого смысла.

Поиск по тексту может выполняться либо путем точного сопоставления строк, либо по произношению. Поскольку данные о произношении не встроены в текст, в компьютерных системах их нужно хранить отдельно. Внедрение поиска аналогично сортировке для японского языка и требует добавления в базу данных дополнительных столбцов для хранения данных пиньинь/йоми. Это может потребоваться как для китайского, так и для японского языка. Дополнительные сложности могут создаваться другими лингвистическими и орфографическими нормами японского языка.

Иногда необходимо, чтобы веб-сайт поддерживал как упрощенное, так и традиционное письмо китайского языка. Если оба типа письма должны отображаться в пределах одного экрана, кодировка Юникод является предпочтительным, если не единственным, решением с технической точки зрения.

Особенностью Тайваня является использование двух языковых стандартов с так называемой проблемой «100-го года». Вторая страна с такой проблемой — Северная Корея. Аналогично проблеме 2000 года, в некоторых приложениях для хранения данных о годе используется только две цифры. Напомним, что на Тайване применяется календарь «республиканской эры» и 2011 г. по нему — это 100-й год, который может интерпретироваться программным обеспечением как 0-й, вызывая тем самым определенные проблемы.

Зачастую в восточноазиатском пользовательском интерфейсе возникают проблемы, связанные со шрифтом и размером в пунктах. Иногда для заданного шрифта в китайском и японском языках размер в пунктах должен быть больше, чем в англоязычной среде, и очень часто этого нельзя добиться простым увеличением размера на один пункт.

Если в диалоговом окне с несколькими вкладками не остается места для расширения текста на японском языке, возможно, потребуется создать новые вкладки и перенести в них часть функциональных возможностей, чтобы освободить место для более длинного перевода на японский язык.

Проект и перевод

Проекты локализации для китайского и японского языков обычно длятся дольше, чем проекты для европейских языков. Много времени занимают тех-

нические работы по учету различных культурных и лингвистических норм. Кроме того, дольше длятся перевод и тестирование.

Поскольку перевод выполняется с учетом региональных и культурных особенностей, в пределах одного языка могут существовать различные варианты, связанные с лексическими отличиями и региональными предпочтениями. Авторам приходилось работать над проектами для японского языка, в которых японской стороне потребовалась более подробная информация и объяснения помимо предоставленных данных в исходных документах. Таким образом, переводчик может создавать более подробный перевод по сравнению с текстом оригинала. Для самого конечного продукта это не проблема, однако, в памяти переводов или результатах синхронизации сопроводительной документации к программному обеспечению могут встречаться пары предложений, которые не соответствуют друг другу в полной мере, а также дополнительный японский текст, у которого нет соответствия в оригинале.

Интернационализация программного обеспечения для китайского и японского языков — это сложный процесс, но при тщательном планировании, анализе, проектировании и выполнении все проблемы преодолимы. В конце концов, успешная интернационализация программного обеспечения для этих двух языков может открыть дополнительные коммерческие возможности.

Фрэнк Лин *— руководитель команды по разработке программного обеспечения для международного и отечественного рынков в компании CareFusion.*

Ангелика Церфасс *— внештатный инструктор и консультант по средствам перевода и связанным с переводами процессам.*