Йошики МИКАМИ

Нагаока, Япония

Профессор Технологического университета г. Нагаоки

Член Исполнительного комитета Всемирной сети в поддержку языкового разнообразия MAAYA

Инициатор межгосударственных проектов «Языковая обсерватория», «Сеть азиатских языковых ресурсов» и «Управление страновыми доменами»

Кацуко ТАНАКА-НАКАХИРА

Нагаока, Япония

Доцент Технологического университета г. Нагаоки

Участник проекта «Языковая обсерватория» и проекта по разработке системы наблюдения за цифровым неравенством в киберпространстве

Оценка языкового разнообразия в Интернете

Введение

Мы хотели бы от лица проекта «Языковая обсерватория» и от лица MAAYA, Всемирной сети с участием многих заинтересованных сторон в поддержку языкового разнообразия, поделиться опытом оценки языкового разнообразия в киберпространстве. Наш доклад состоит из трех частей. Он начинается с краткого описания проекта «Языковая обсерватория», далее представлены некоторые результаты нашего исследования и в конце обсуждается интерпретация полученных результатов.

Что такое «Языковая обсерватория»

Начнем с вопроса о том, что такое «Языковая обсерватория». Это название было придумано по аналогии с астрономической обсерваторией. В астрономической обсерватории наблюдают за звездами на небе, а в языковой обсерватории ведется наблюдение за языками в киберпространстве. Мы знаем, что существует более пяти тысяч звезд, которые можно различить на небе невооруженным гла-

зом. Количество языков, на которых говорят на Земле, превышает шесть тысяч. Но лишь в отношении некоторых реализованы преимущества информационных технологий. Поисковые системы могут работать только с ограниченным числом языков; основные компьютерные платформы до сих пор поддерживают лишь немногие из них. Осознав это, мы решили организовать виртуальную обсерваторию, чтобы получить подлинное представление о ситуации.

В языковой обсерватории два основных технических компонента. Первый – это поисковый робот, обходящий веб-страницы. Второй – идентификатор языков, который автоматически устанавливает реквизиты языка для отобранной страницы.

Говоря об «идентификации языка», мы имеем в виду не только языки, но и системы письма, а также кодировки страниц. В настоящее время наша обсерватория способна различать более трехсот языков.

Проект был начат четыре года назад, и его запуск совпал с принятием Рекомендации ЮНЕСКО по киберпространству[17] (рекомендация была принята всего за несколько месяцев до начала проекта). Измерение языкового разнообразия в киберпространстве – одна из основных проблем, затронутых в этой рекомендации. В Международный день родного языка в 2004 г. мы организовали первый семинар и пригласили сотрудника ЮНЕСКО на церемонию, посвященную запуску проекта. Это событие освещалось на новостном сайте ЮНЕСКО.

Финансирование, полученное от Японского агентства по науке и технике (JST), позволило нам создать парк служебных компьютеров. Кроме того, успех проекта определяется технической поддержкой, которую оказали наши зарубежные сотрудники. Нам очень помогла мощная программа-обходчик UbiCrawler, разработанная группой специалистов Миланского университета в Италии.

Некоторые результаты исследования

Теперь мы хотели бы коротко изложить результаты исследования.

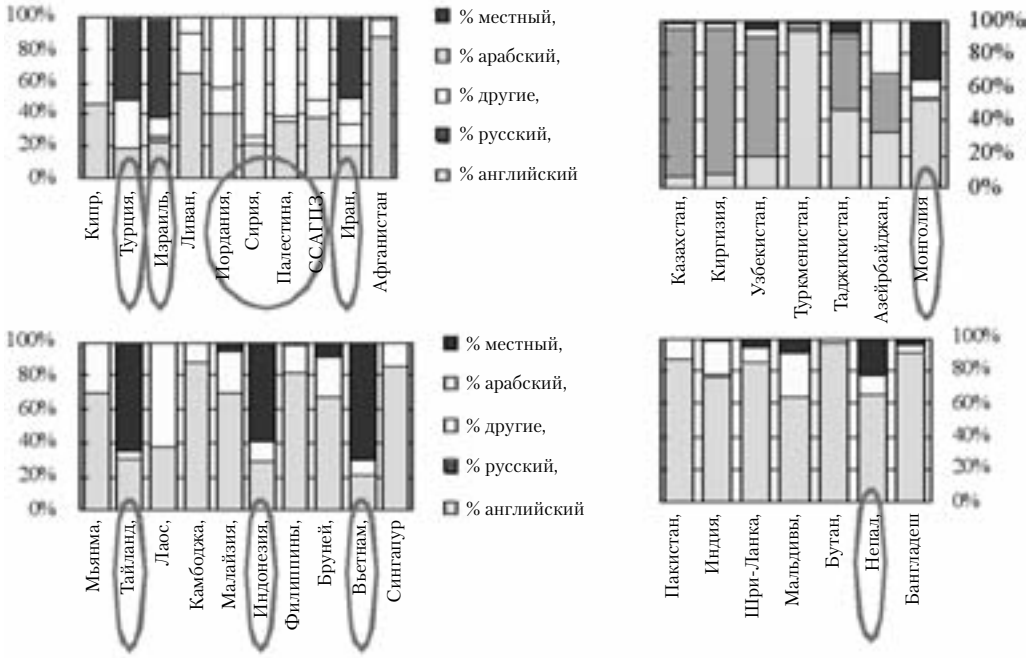
Всю Сеть просмотреть невозможно из-за ее чрезвычайно большого объема. По некоторым оценкам, ее размер составляет десятки миллиардов страниц. Поэтому мы решили сконцентрировать усилия на доменах стран Азии и Аф-

[17] Рекомендация о развитии и использовании многоязычия и всеобщем доступе к киберпространству, принята Генеральной конференцией ЮНЕСКО на 32-й сессии 15 октября 2003 г. (Прим. ред.)

рики. На рисунке 1 показан состав языков веб-страниц 34 стран Азии[18] (не включены Китай, Корея и Япония). Как можно увидеть, местные языки занимают господствующее положение лишь в доменах немногих стран, а именно: Турции, Израиля, Ирана, Таиланда, Индонезии и Вьетнама, а также нескольких арабоговорящих государств Ближнего Востока. Голубой цвет соответствует доле веб-страниц, написанных на английском языке, и неудивительно, что в некоторых национальных доменных зонах большая часть страниц (порой свыше 90 %) создана именно на английском.

Рисунок 1

Результаты исследования. Азия



В таблице 1 показана первая десятка местных языков Азии с соответствующим числом носителей и количеством страниц, найденных нами в Сети. На первом месте – иврит, за ним следуют тайский, турецкий, вьетнамский, арабский и т.д.

[18] Здесь и далее используется классификация стран по регионам, принятая ЮНЕСКО (прим. ред.).

Таблица 1

Примерное количество страниц 10 наиболее употребимых азиатских языков

Язык	Система письма	Количество говорящих	Количество страниц
Иврит	Иврит	4,612,000	11,957,314
Тайский	Тайское письмо	21,000,000	7,752,785
Турецкий	Латиница	59,000,000	3,959,328
Вьетнамский	Латиница	66,897,000	2,006,469
Арабский	Арабское письмо	280,000,000	1,671,122
Татарский	Латиница	7,000,000	1,575,442
Фарси	Латиница	33,000,000	1,293,880
Яванский	Латиница	75,000,000	1,267,981
Индонезийский	Латиница	140,000,000	866,238
Малайский	Латиница	17,600,000	432,784

Примечание: китайский, корейский и японский домены не учитывались.
По состоянию на октябрь 2006 г.

На Африканском континенте ситуация намного хуже. Более шестидесяти доменов стран объединены в три языковые группы: англоговорящая (Содружество наций), франкоговорящая и арабоговорящая. И опять вполне естественно, что в каждом из них большую часть занимают английский и французский языки. При этом во всех группах местные африканские языки присутствуют в почти пренебрежимо малом количестве.

В таблице 2 показана первая десятка местных африканских языков вместе с соответствующими регионами и количеством страниц, найденных нами в Сети. На первом месте – малагасийский язык, официальный язык Мадагаскара, за ним следуют суахили, африкаанс, крио, киньярванда и т.д. Но местные африканские языки представлены в намного меньшей степени, чем азиатские. Это в лучшем случае около нескольких тысяч.

Таблица 2

Примерное количество страниц на 10 наиболее употребимых африканских языках

Язык	Система письма	Регион, в котором используется язык	Количество страниц
Малагасийский	Латиница	Мадагаскар	5,382
Суахили	Латиница	Танзания	5,170
Африкаанс	Латиница	ЮАР, Намибия	1,775
Крио	Латиница	Гамбия, Сьерра-Леоне	1,575
Киньярванда	Латиница	Руанда	1,059
Шона	Латиница	Зимбабве, Мозамбик	538
Сомали	Латиница	Сомали	396
Сисвати	Латиница	Свазиленд	335
Ошиванбо	Латиница	Намибия, Ангола	264
Рунди	Латиница	Бурунди	252

Примечание: ЮАР не включена.
По состоянию на октябрь 2006 г.

Как интерпретировать полученные результаты?

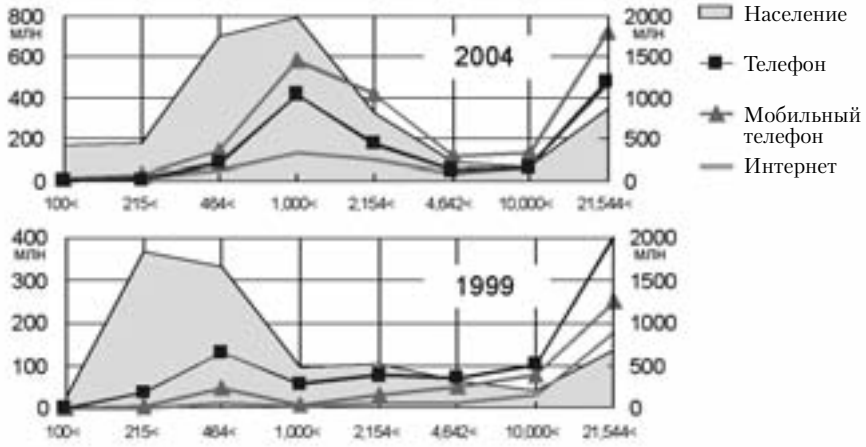
Приведенные данные свидетельствуют о существовании серьезного разрыва между языками. Его можно назвать «языковым цифровым разрывом». В данном разделе будут интерпретироваться полученные результаты.

Во-первых, хотелось бы представить некоторые данные в экономическом контексте. На рисунке 2 показано распределение населения и уровень доступа к средствам связи в странах с разным уровнем дохода. Справа располагаются страны с высоким доходом, а слева – страны с низким доходом; два графика относятся к 1999 и 2004 гг. За эти пять лет заметен значительный прогресс в телефонии, в особенности в отношении мобильной связи.

Рисунок 2

Как интерпретировать полученные результаты?

Экономический аспект

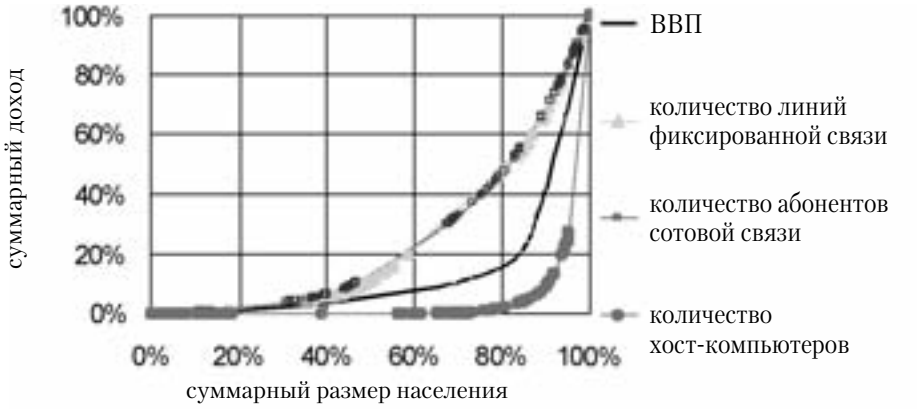


Ресурс: ITU Statistics ВВП на душу населения, в долл.США

Экономисты, говоря о неравенстве доходов, постоянно ссылаются на коэффициент Джини. Если отложить суммарный размер населения по горизонтальной оси, а суммарный доход – по вертикальной оси, получаем кривую Лоренца. Чем больше изгиб кривой, тем выше неравенство доходов. Это понятие было заимствовано для измерения разрыва в киберпространстве.

На рисунке 3 показаны кривые Лоренца по четырем показателям: ВВП, количество телефонов фиксированной связи, число абонентов мобильной телефонной связи и число хост-компьютеров, подсоединенных к Интернету. Коэффициент Джини для телефонии – 0,73. Он ниже соответствующего показателя для ВВП. Иными словами, неравенство, обнаруживаемое в доступе к телефонной связи, сейчас ниже, чем неравенство доходов. Но неравенство хост-компьютеров все же выше, чем неравенство доходов. Данный анализ ясно показывает, что ситуация с телефонией улучшилась, но с доступом в Интернет – все еще нет.

Рисунок 3



Коэффициент Джини: Телефония 0.51 < ВВП 0.73 < Интернет 0.91

Мы хотели бы также ввести один фактор с технической точки зрения. Это карта мира, окрашенная в разные цвета согласно системам письма, используемым для отображения родных языков регионов (рис. 4).

Рисунок 4

Всемирная карта систем письма

На карте большая часть мира окрашена в два немного отличающихся оттенка желтого цвета, которые соответствуют латинскому и кириллическому алфавитам. Лиловый цвет соответствует арабскому языку и используется для Северной Африки и Среднего Востока. Но Азиатский континент выглядит как лоскутное одеяло, составленное из множества цветных кусков, что создает в регионе особые технические трудности.

Эти трудности иллюстрирует один старинный документ. В письме, написанном четыреста лет назад иезуитским монахом, говорится, что в то время сложность задачи литья более шестисот литерных матриц служила препятствием к печатанию книги на местном языке.

Еще один пример старопечатной книги – отпечатанная в Маниле «Doctrina Christiana» – дает представление о том, что происходило в наихудшем случае. Книга была напечатана в трех вариантах: на тагальском языке тагальским письмом, на тагальском языке латинским шрифтом и на испанском языке латинским шрифтом. Уже в первые сто лет, прошедших с момента первой печати, тагальский шрифт был полностью утрачен.

У нас есть коллекция пишущих машинок, адаптированных для печати на разных языках, в том числе тамильская, бенгальская, сингальская, английская, хинди, корейская, мьянманская и тайская. Удивительно, что все пишущие машинки имеют аналогичную форму и почти одно и то же количество клавиш, несмотря на то что количество букв в этих языках совершенно различно. Почему? При адаптации пишущих машинок широко использовались компромиссные пропуски букв и сложные наложения специальных наборных знаков, причем они не всегда были удобны для местных пользователей.

Как мы объяснили на трех примерах, с начала эры печати именно трудности локализации на тот или иной язык были основным препятствием на пути распространения новых информационных технологий. Очевидно, в эпоху компьютеров центральным фактором в работе по локализации являются проблемы кодирования символов.

Это предположение подтверждается таблицей 3. Первая пятерка языков – это те языки, которые широко представлены в киберпространстве, и для них всех используется жестко установленный стандарт кодирования соответствующих систем письма. Последние четыре языка в таблице, напротив, страдают от хаоса в кодировании и довольно ограниченной представленности в Сети.

Таблица 3

Хаос в кодировании приводит к замедлению в локализации

Язык	Стандарт кодирования и его распространение	Встречающиеся примеры других кодировок
Турецкий	ISO 8859 (99.5 %)	
Иврит	ISO 8859 (87.7 %)	
Вьетнамский	UTF-8 (96.4 %)	TCVN, VIQR, VPS
Тайский	TIS 620 (97.3 %)	
Монгольский	UTF-8 (95.5 %)	Latin-Cyrillic
Синхала	UTF-8 (44.5 %)	Metta, Kaputa и др.
Телугу	UTF-8 (16.6 %)	Shree, TLH и др.
Тамильский	UTF-8 (14.9 %)	Amudham, Kumudam, Shree, Vikatan и др.
Бирманский	UTF-8 (0.7 %)	WinResearcher и др.

Примечание: Местные запатентованные кодировки показаны в этой таблице по названиям шрифтов по состоянию на июнь 2006 г.

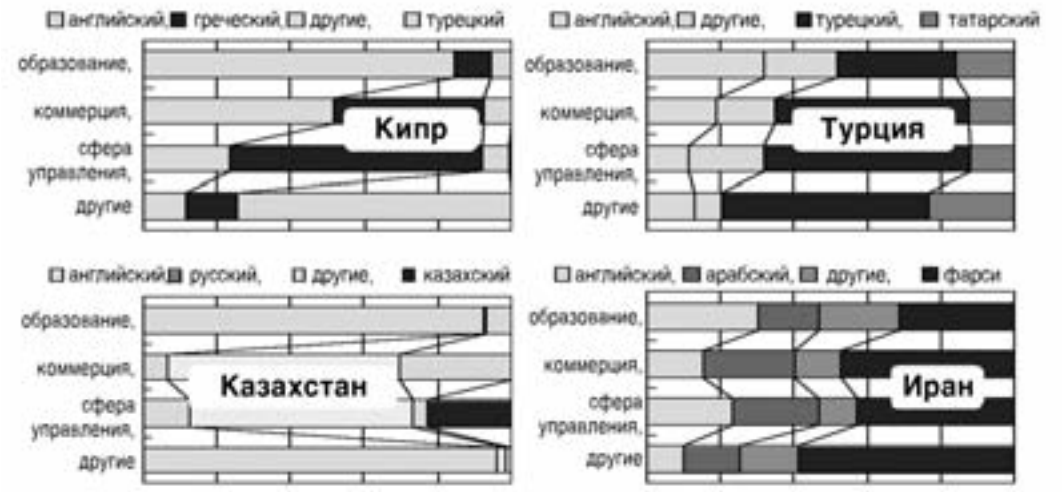
Наконец, мы хотели бы представить социокультурный аспект. Языковая деятельность людей происходит в разных сферах. В данном случае нами заимствована схема анализа из документа Европейского Союза. В документе выделены четыре сферы: личная, публичная или государственная, профессиональная или деловая и образовательная.

В одноязычном обществе типа Японии во всех четырех сферах используется один язык. Но в многоязычных обществах в разных сферах функционируют разные языки. Например, языки международного общения, такие как английский или французский, работают в образовательной и профессиональной сферах, в то время как в повседневном общении используются местные языки вплоть до диалектов меньшинств. Если проанализировать состав языков в сферах вторичного уровня каждой страны, результат будет иллюстрировать картину специализации каждого языка.

На рисунке 5 представлены результаты выборочного обследования по четырем странам: Кипру, Турции, Казахстану и Ирану. Мы наблюдаем высокий процент использования английского языка в образовательных сферах всех четырех стран, но в личном общении преобладают официальные или местные языки в большинстве стран.

Рисунок 5

Специализация языка: анализ вторичного уровня



Заключение

Теперь хотелось бы подвести итог сказанного.

Как показано во второй части доклада, языковой цифровой разрыв наблюдается в доменах стран Азии и Африки. Мы предположили, что в основе этого разрыва лежат три фактора, соответствующие экономическому, техническому и социокультурному аспектам. С экономической точки зрения один из факторов – это меньшие возможности доступа в странах с более низким доходом. С точки зрения технического аспекта, как уже упоминалось, это особые трудности локализации. Они серьезны прежде всего для тех, кто не пользуется латинским письмом. С точки зрения социокультурного аспекта это меньшее присутствие местных языков в образовательной и профессиональной сферах. Это означает, что требуется расширение прав и возможностей местных языков.

Что касается будущего «Языковой обсерватории», то мы видим два направления работы. Первое – создание на базе нашей технологической инфраструктуры поисковых систем по отдельным языкам. Для таких систем ключевыми элементами являются также идентификация языков и поисковый робот. Второе направление – построение сети языковых обсерваторий в мировом масштабе. Надеемся, что найдутся люди, которые согласятся участвовать в этом направлении деятельности и совместно работать в будущем.