
Языковое разнообразие в Интернете

Джон Паолилло, Институт информатики,
Университет Индианы

Прошло уже более 20 лет с момента появления Интернета в англоязычном мире, а представленность различных языков в этом пространстве по-прежнему сильно смещена в направлении преобладания английского языка. Английский остается доминирующим языком в Интернете, а это значит, что языки, на которых говорит множество людей, представлены либо плохо, либо не представлены совсем. В какой степени эта ситуация отражает предпочтение, отдаваемое английскому языку, и отказ от других языков? В данном докладе¹⁷ мы рассмотрим этот вопрос с позиций этической системы Фридмана и Ниссенбаума (Friedman and Nissenbaum, 1997), разработанной для оценки сдвига в компьютерных системах и сравнения полученных данных со статусом мировых языков в Интернете. Эта система поможет нам интерпретировать возможные причины и способы корректировки потенциальных сдвигов. Мы также обратим внимание на звучащее в настоящее время недовольство статусом языков в Интернете и рассмотрим эту ситуацию

17 Помощь в написании настоящего доклада оказали: Elijah Wright и Hong Zhang (Университет Индианы), S., Baskaran, G.V., Ramanan, S.V., Rameshkumar, S., Shoba Nair, I., Vinoshbaru James, Viswanathan (Университет Св. Анны, Ченнай, Индия). Полная версия доклада представлена на <http://ella.slis.indiana.edu/paolillo/paolillo.divesity041027.pdf>

в рамках указанной системы, которая требует от нас анализа не только распределения и использования языков в Интернете, но и анализа общественных институтов, осуществляющих управление развитием Интернета, которые, по выражению Фридмана и Ниссенбаума, могут вызвать «возникающие сдвиги». Наконец, мы рассмотрим вопросы языковых сдвигов в технических системах Интернета.

Сдвиг, многоязычие и компьютерные системы

Цифровой разрыв, т. е. неравное распределение доступа к источникам и сервисам цифровой информации, стал одним из ключевых моментов в наше время цифровой информации. Правительства, международные агентства, группы граждан, корпорации и другие объединения стремятся овладеть преимуществами, которые сулят обещания низкой стоимости и мгновенного доступа к информации за счет перевода многих коммуникационных функций на сетевые компьютерные носители. Однако, если традиционные социальные барьеры, такие как социально-экономический статус, образование, этническая принадлежность, пол и пр. затрудняют доступ к цифровой информации, то политика должна быть направлена на выравнивание доступа для реализации обещанных преимуществ.

Вопрос статуса мировых языков можно рассматривать в свете цифрового контента. Некоторые языки имеют огромные объемы доступного цифрового контента. Пользователи Интернета, говорящие, читающие и пишущие на таких языках, имеют намного меньше проблем, связанных с доступом и обменом полезной информацией, чем люди, говорящие на языках, хуже представленных в информационных системах. Анализ этой ситуации наводит на мысль о том, что, сдвиг, возможно, создается самими цифровыми информационными системами, их конфигурацией или использованием. Стало ли языковое различие барьером на пути доступа к информации, который предоставляет несправедливые преимущества одним и ставит в невыгодное положение других? Анализ подобных вопросов этического и морального порядка требует создания специальной системы.

ЮНЕСКО и культурное разнообразие

В 2001 г. государства – члены ЮНЕСКО приняли Всеобщую декларацию о культурном разнообразии¹⁸. Статья 6 «К культурному разнообразию, доступному для всех» гласит:

«Обеспечивая свободное распространение идей словесным и изобразительным путем, следует добиваться, чтобы все культуры могли быть объектом самовыражения и распространения. Залогом культурного разнообразия является свобода выражения мнения, плюрализм средств информации, многоязычие, равный доступ к возможностям для художественного творчества, к научно-техническим знаниям, в том числе в цифровой форме, и обеспечение всем культурам доступа к средствам выражения и распространения идей».

Опираясь на данную Декларацию, ЮНЕСКО недвусмысленно поддерживает предоставление всем языковым и культурным группам равного доступа к цифровой информации, как в сфере производства, так и в сфере потребления. Декларация конкретизирует эту позицию, перечисляя направления действий, которые необходимы для претворения ее в жизнь. Три из них имеют непосредственное отношение к вопросам цифровых носителей и информационных технологий:

9. Стимулирование «цифровой грамотности» и обеспечение более высокого уровня владения новыми информационными и коммуникационными технологиями, которые следует рассматривать и как учебную дисциплину, и как педагогические средства, способные повысить эффективность образовательных услуг.
10. Продвижение языкового разнообразия в киберпространстве и стимулирование всеобщего доступа через глобальную сеть к информации, являющейся общественным достоянием.
11. Преодоление цифрового разрыва в тесном сотрудничестве с соответствующими организациями системы ООН путем по-

18 <http://unesdoc.UNESCO.org/images/0012/001271/127160m.pdf>

Прим. ред. рус. изд.: Всеобщая декларация о культурном разнообразии на русском языке размещена на веб-сайте Российского комитета Программы ЮНЕСКО «Информация для всех»: <http://www.ifapcom.ru/ru/387/425/>

ощрения доступа развивающихся стран к новым технологиям и облегчения распространения местного контента в цифровом формате и доступа этих стран к образовательным, культурным и научным цифровым ресурсам, доступным во всем мире (ЮНЕСКО, 2001, стр. 6).

Эти принципы и направления действий определяют критерии оценки признаков информационного общества в терминах этики и задач развития такого общества. Однако они не предоставляют достаточно глубокого анализа возможных причин любых сдвигов, которые могут проявиться. Без такого анализа достаточно трудно дать правильные рекомендации по необходимым действиям.

Так, например, электронные библиотеки не встретили одобрения у народа маори в Новой Зеландии. Внимательное изучение показало, что это не просто вопрос цифровой грамотности, а отражение культурных традиций, препятствующих успешному использованию электронных ресурсов. Выяснилось, что метафора библиотеки относится к форме организаций, принадлежащих «Pakeha» (белым, выходцам из Западной Европы), предполагающих доступ к информации, чуждой для народа маори (Dunker, 2002). Ключевое место конфликта для маори – открытый доступ к информации, которая традиционно защищается в их культуре. Такой информацией являются для них генеалогические данные. Концепция библиотек, применяющих принцип открытого доступа к информации независимо от ее содержания, вступила в конфликт с ценностями маори. Следовательно, модель информационного доступа для электронных библиотек необходимо пересмотреть для того, чтобы создать электронную библиотеку, признанную народом маори¹⁹.

Этическая система

Фридман и Ниссенбаум (Friedman and Nissenbaum, 1995, 1997) создали полезную основу для анализа сдвигов в компьютерных системах, помогающую сосредоточить внимание на причинах существующих предпочтений. Они определили 3 основные категории сдвигов: исторические, технические и возникающие. Исторические сдвиги берут начало

19 Эта ситуация напоминает проблемы, связанные со случайным опубликованием в Интернете личных медицинских карт людей.

в общественных институтах, практиках и отношениях, и существуют независимо от компьютерных систем. Технические сдвиги возникают тогда, когда технические свойства используемых систем не учитывают всех особенностей той сферы, в которой они применяются. Возникающие сдвиги появляются в процессе работы реальных пользователей; они не являются осознанной частью проекта или системы, или неотъемлемой частью социального контекста, а возникают из взаимодействия того и другого в определенной ситуации.

Проиллюстрируем все три формы сдвигов на языковых примерах. Исторические сдвиги проявляются тогда, когда правительство, отрасль промышленности или мощная корпорация отказываются предоставлять носителям одного или нескольких языков информацию, технологии или продукты. Так, в середине 1990-х годов Microsoft Inc. отказалась сделать версии своих продуктов, которые могли бы работать с алфавитами, отличными от латинского, несмотря на наличие уже имеющихся на рынке технологических решений, таких как WorldScript, созданный Apple Computer Inc. Microsoft Inc. объявила, что отказывается сделать это потому, что рынки языков, не использующих латиницу, слишком малы, чтобы оправдать создание новой версии их продукта; так, мы видим, что за этим примером исторического сдвига стояла экономическая мотивация²⁰.

Технические сдвиги возникают в системах кодирования текстов, таких как Юникод UTF-8, и приводят к тому, что текст не на латинице требует в 2–3 раза больше места, чем аналогичный текст на латинице. В этом случае мотивация заключается в вопросах совместимости старых систем, основанных на латинице, с более молодыми системами, использующими Юникод.

Наконец, возникающие сдвиги появляются тогда, когда компьютерные системы, разработанные для одной цели, используются для другой цели, например, когда система электронных библиотек, разработанная для городских условий, в которых проживает белое население Новой Зеландии, воспринимается с неодобрением народом маори, проживающим в сельской местности.

20 За прошедшее время Microsoft изменила свою позицию и создала версии своих продуктов для рынков других языков.

Каждый из трех типов сдвигов требует отдельного подхода. Исторические сдвиги следует рассматривать в свете образовательных, юридических и институциональных ресурсов стран, отраслей промышленности или корпораций. Технические сдвиги следует рассматривать при разработке основополагающих принципов создаваемых компьютерных систем. Возникающие сдвиги следует рассматривать через образование и создание проектов, основанных на наблюдениях за реальным использованием компьютерных систем.

Поскольку развитие Интернета предполагает взаимодействие технологий, предварительных условий, целей, отраслей промышленности и действующих лиц, все три типа сдвигов отражаются на развитии языков в Интернете, в разных местах и в разное время.

Интернационализация и Интернет: популярные концепции

Дискуссии в СМИ о потенциальном языковом сдвиге в Интернете развиваются в двух противоположных направлениях. Вассерман (Wasserman) так описывает эту ситуацию:

«Поскольку Интернет способствует... пониманию мира как взаимосвязанной и взаимозависимой среды, его можно рассматривать как один из факторов ускорения глобализации... Поскольку глобализация рассматривается как сила, источником которой является так называемый экономически развитый мир, некоторые критики предсказывают разрушение отличительных местных и культурных особенностей в миноритарных странах и сообществах. С другой стороны, некоторые критики говорят, что глобальные и локальные силы взаимодействуют в процессе глобализации, превращая ее в многосторонний процесс, приносящий пользу и даже стимулирующий развитие местных культур и языков. (Wasserman, 2002:2)».

Те, кто придерживается первого взгляда, выступают защитниками прав меньшинств, а те, кто придерживается второго взгляда, выступают сторонниками новых сетевых информационных технологий. Первый взгляд – это до некоторой степени реакция на быстрые и далеко идущие изменения, вызванные распространением Интернета, в то время как второй взгляд был с самого начала в значительной степени спровоцирован создателями этой технологии.

Достаточно просто найти популярные отчеты, отражающие работу команд инженеров над ранее существовавшей сетью ARPANET (первой компьютерной сетью) как над идеализированной децентрализованной демократической организацией (см. Hafner and Lyon, 1996), или над сетью Whole Earth 'Lectronic Link (известной также как WELL), способствующей созданию виртуальных Интернет-сообществ (Rheingold, 2000). От этого взгляда совсем недалеко до теории, которая считает языковое доминирование одной из форм неравенства, которую технология Интернета сможет быстро устранить. Во-первых (как говорят сторонники этой теории), Интернет – это глобальная и децентрализованная сеть; никакой пользователь или группа пользователей не может осуществлять иерархический контроль над другим пользователем или группой пользователей, потому что Интернет допускает полную свободу общения. Как следствие, каждый желающий может пользоваться любым языком, если есть еще хотя бы один человек, который также хочет пользоваться этим языком. Во-вторых, рост числа неанглоязычных пользователей, в особенности пользователей, пишущих и читающих на китайском языке, должен, по прогнозам, превысить текущий показатель роста англоязычных пользователей. Иными словами, в конце концов, английский язык не будет доминировать в Интернете, потому что на других языках будет говорить гораздо больше людей. Вопрос о том, какой язык является доминирующим в сети, на самом деле относится к разряду вопросов демографического распределения. И, наконец, сторонники Интернета говорят, что такие технические возможности Интернета, как Юникод для многоязычных текстов и системы типа BabelFish для моментального перевода веб-документов, способны разрешить любые проблемы, связанные с использованием информации в Интернете, с которыми могут столкнуться люди, говорящие на других языках. В частности, эта перспектива в значительной степени характеризует позицию, отраженную в издании «Культурное и языковое разнообразие в информационном обществе», подготовленном ЮНЕСКО для Всемирного саммита по информационному обществу (UNESCO, 2003)*.

* *Прим. ред. рус. изд.:* эта книга издана на русском языке Российским комитетом Программы ЮНЕСКО «Информация для всех» и Российской национальной библиотекой: <http://www.ifapcom.ru/365/432/>

На каждый из этих аргументов есть контраргумент, представляющий альтернативный взгляд, согласно которому английский язык и, в меньшей степени, другие европейские языки являются доминирующими в коммуникации посредством Интернета. Причины этого явления частично социальные, частично технические. Во-первых, приводится аргумент в пользу того, что Интернет использует телекоммуникационную инфраструктуру, экономическое доминирование в которой принадлежит компаниям США. Географическим центром связи в глобальной телекоммуникационной сети являются США, так что любое действие, способствующее этой ситуации, будет непропорционально выгодно для США за счет более низкой стоимости связи и большого числа абонентов. Во-вторых, несмотря на наметившиеся в последнее время тенденции, англоязычные пользователи по-прежнему составляют самую многочисленную группу пользователей Интернета. В-третьих, представительство англоязычных пользователей Интернета диспропорционально по отношению к представительству народов, говорящих на других языках. Наконец, большинство технологий, используемых в Интернете, лучше всего адаптировано к английскому языку. Интерфейсы для нелатинских алфавитов громоздки, а для ряда языков не существуют вообще. Даже такие системы, как Юникод, инкорпорируют технические сдвиги, отражающие предпочтение английского языка, а системы перевода недостаточно совершенны, чтобы соответствовать требуемому уровню²¹.

Эти точки зрения отличаются подходами, которые составляют основу трех типов сдвигов, определенных Фридманом и Ниссенбаумом (Friedman and Nissenbaum, 1997). Языковая демография пользователей Интернета заставляет задуматься об исторических сдвигах. Вопрос доступности технических возможностей для разных языков поднимает тему технических сдвигов. Кроме того, вопросы децентрализации вместо существующего де-факто централизованного контроля заставляют говорить о возникающих сдвигах в системе, которая вышла в своем развитии за рамки первоначальных, национальных границ.

21 Варианты этих двух точек зрения и их отношение к аналогичным последствиям глобализации обсуждаются в Block (2004).

Несмотря на различные мнения и периодически появляющиеся жаркие дебаты, мы ощущаем явный дефицит эмпирических исследований, которые непосредственно занимались бы вопросами исторических, технических и возникающих сдвигов в Интернете. Причина этого частично заключается в том, что Интернет огромен по масштабам и подвержен быстрым изменениям. Оба эти обстоятельства сильно затрудняют получение надежных данных. И, хотя лингвистические обзоры периодически проводятся маркетинговыми организациями, такими как Jupiter Research (<http://jupiterresearch.com>) и Global Reach (<http://www.greach.com/>), ценность их данных по оценке вопросов языковых сдвигов сомнительна по причине наличия экономических интересов у специалистов по маркетингу и их клиентов. Более того, надежный крупномасштабный анализ онлайн-многоязычия будет стоить дорого и потому недоступен научным организациям, имеющим ограниченный бюджет или вообще не получающим финансирование.

Источники исторических сдвигов

Исторические сдвиги затрагивают общественные институты, практику и отношения, независимо от технологий. В числе источников такого типа сдвигов – историческое распределение населения по языкам, экономические меры, способствующие процветанию более многочисленных языков, и институциональная политика государств. Что касается языкового разнообразия в Интернете, то исторические сдвиги имеют место в отношении правительств, институтов и компаний к людям, говорящим на разных языках, при реализации политики в области информационных технологий. Понять эти сдвиги непросто, но чрезвычайно важно, и, поскольку Интернет – явление глобальное, понимать их нужно в контексте глобального языкового разнообразия.

Глобальное языковое разнообразие

Логически построенное обсуждение языкового разнообразия в глобальном или региональном масштабе требует использования количественного показателя разнообразия. К сожалению, количественное измерение языкового разнообразия редко используется в современных языковых исследованиях, а широко используемой общепринятой меры не существует. Существующие критерии по числу языков или языко-

вых групп представляются несколько упрощенными (Barrera-Brassols and Zenck, 2002; Smith, 2001). Более сложные показатели разнообразия предлагались (Greenberg, 1951; Lieberman, 1964), но не получили статистического обоснования и перестали использоваться. Подход, о котором мы расскажем в настоящей статье, основан на методе Неттла (Nettle, 1999) и заключается в применении меры вариативности в качестве показателя разнообразия.

Удовлетворительный индекс языкового разнообразия должен учитывать несколько факторов. Во-первых, он должен отражать некий объект анализа – страну, континент или Интернет. Во-вторых, языковое разнообразие должно учитывать вероятность нахождения людей, говорящих на любом выбранном языке. Он должен иметь естественный равный нулю минимум для абсолютно однородного населения и не иметь фиксированного максимума. Чем больше языков, тем больше должно быть значение индекса, но по мере уменьшения доли языковой группы ее вклад в разнообразие тоже должен уменьшаться. Страны с большим числом языковых групп примерно одинаковой численности (например, Танзания; Mafu, 2004) покажут относительно высокое языковое разнообразие, в то время как страны с примерно таким же числом языков, но с одним или двумя доминирующими языками (такие как США), продемонстрируют относительно низкое языковое разнообразие. Мера, обладающая такими характеристиками, – информационно-теоретическое образование под названием «энтропия», на которой мы строим нашу меру языкового разнообразия. В статистических терминах энтропия есть мера вариативности. Энтропия складывается из известного количества населения страны, говорящего на каждом языке, умноженного на его натуральный логарифм и суммированного со всеми записями для данной единицы – страны или региона. Окончательная величина индекса будет в 2 раза больше этой суммы.

Таблица 1 и рисунок 1 представляют данные по этой основанной на энтропии мере для разных регионов мира на базе 7639 языковых групп, представленных в Ethnologue (www.ethnologue.com) в порядке от самого низкого к самому высокому уровню языкового разнообразия. Данные США, родины Интернета, были размещены в первом ряду для сравнения. Регионы, известные своим языковым разнообразием (Африка, Океания), показывают наивысшее языковое разнообразие, а регионы с крупными национальными языками (Восточная Азия, Северная Америка) показывают самое низкое языковое разнообразие. Два по-

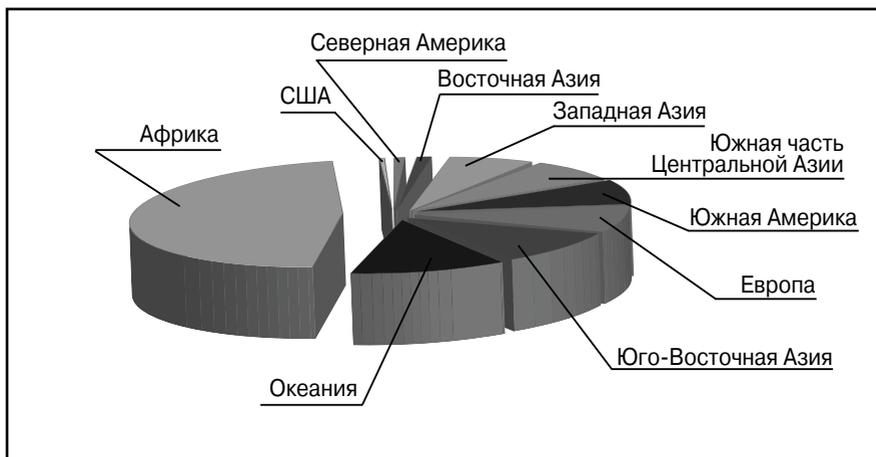
следних региона особенно важны для понимания языкового разнообразия в Интернете: США и Китай, вероятно, представляют собой двух крупнейших игроков в Интернете (по некоторым прогнозам в ближайшие несколько лет Китай обойдет США по числу пользователей); ни одна из этих стран не проявляет большего языкового разнообразия в сравнении со странами Океании или Африки. До тех пор, пока эти страны доминируют в Интернете (или, если говорить шире, в дискуссии по языковому разнообразию в Сети), Интернет не может отражать языковое разнообразие мира.

Таблица 1. Показатели индекса языкового разнообразия по регионам

Регион	Языки	Индекс разнообразия	Доля от общего числа языков в мире
США	170	0,7809	0,0020
Северная Америка (вкл. США)	248	3,3843	0,0086
Восточная Азия	200	4,4514	0,0112
Западная Азия	159	26,1539	0,0659
Южная часть Центральной Азии	661	29,8093	0,0752
Южная Америка	930	30,5007	0,0769
Европа	364	32,4369	0,0818
Юго-Восточная Азия	1317	37,6615	0,0949
Океания	1322	46,5653	0,1174
Африка	2390	185,6836	0,4681

Источник: Ethnologue.

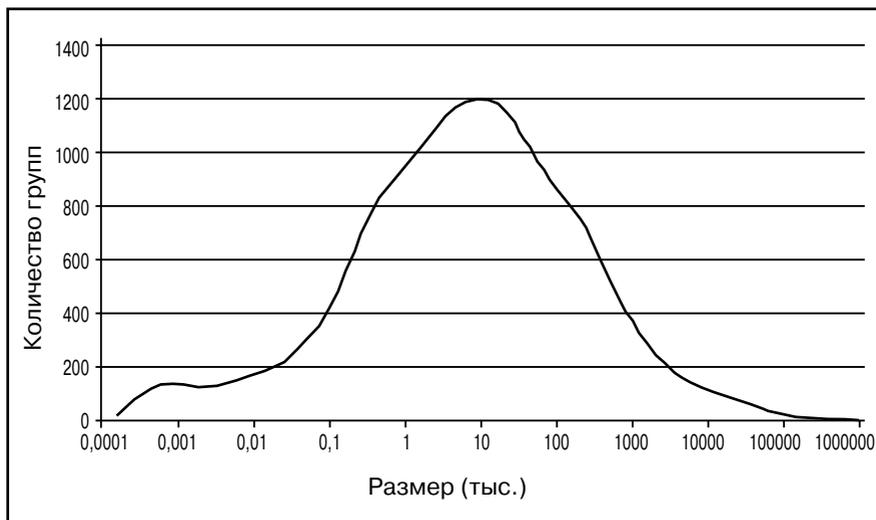
Рис. 1. Индекс языкового разнообразия по регионам



Источник: Ethnologue.

Эволюция языкового разнообразия

Общее представление о языковом разнообразии можно получить в результате изучения размеров языковых популяций. На рис. 2 показано количество языковых групп разной численности, цифры приводятся по данным Ethnologue. Горизонтальная ось представлена в логарифмическом масштабе, что означает, что колоколовидная кривая отражает логарифмически нормальное распределение (Grimes, 1986). Типичный размер языковой группы – несколько десятков тысяч человек, что примерно равно численности населения небольшого города. Языки с сотнями миллионов носителей, такие как английский, китайский, испанский, французский и пр., являются несколько нетипичными, так же как и более мелкие языковые группы с населением в сотни человек. С бытовой точки зрения ситуация выглядит иначе: почти половина людей на Земле говорит на языке, на котором говорят еще сотни миллионов других людей.

Рис. 2. Размеры языковых групп

Источник: Ethnologue, ПРООН.

Глобальное и региональное языковое разнообразие – явление не статическое, оно имеет тенденцию развития во времени. Это разнообразие находится под влиянием социально-исторических событий, таких как массовая миграция, колонизация, войны, эпидемии и пр. Глобальное языковое разнообразие в настоящее время сокращается, и процесс этот идет уже давно. Для лингвистов, изучающих разнообразие человеческой речи, это кризисная ситуация. Вымирание сотен языков за короткий исторический период означает, что большие массивы знаний о способности человека говорить пропадают навсегда, а с ними пропадают литература, история и культура людей, на этих языках говоривших. Для сообществ, чьи языки, история и культура исчезают, ситуация приравнивается к катастрофе.

По некоторым оценкам почти половина языков в мире исчезнет к концу 2050 г. (Dalby, 2003; Krauss, 1992; Nettle and Romaine, 2000). Когда языковое разнообразие перестанет существовать вследствие исчезновения малых языковых групп, возрастет доля населения, относящаяся к большим языковым группам.

Потеря языкового разнообразия не ограничивается каким-то одним регионом мира: языки исчезали в большом количестве в Европе по причине расцвета национальных государств; в Северной и Южной Америке и Австралии драматические потери шли следом за европейской колонизацией и продолжаются до сих пор; на островах Тихого океана и в Индонезии английский и индонезийский вытесняют языки аборигенов; в Азии крупные языки в Китае, Японии, Индии и России веками расширяли сферу своего влияния за счет других языков (Crystal, 2000; Muhlhausler, 1996).

Некоторые причины вымирания языков очевидны. Так, случайная или намеренная ликвидация группы людей может привести к смерти языка (Wurm, 1991). Большая часть языкового разнообразия Северной Америки была потеряна именно таким путем: войны с европейскими поселенцами и болезни, распространявшиеся в результате контактов с европейцами, приводили к сокращению аборигенного населения до такой степени, что некому было поддерживать их родные языки. Другие причины вымирания языков менее очевидны, особенно когда они связаны с изменениями в экологии культуры.

Глобальное языковое разнообразие и Интернет

Низкий уровень языкового разнообразия, характерный для Северной Америки, Латинской Америки и Карибского бассейна, Европы и Восточной Азии, способствует предоставлению доступа к Интернету с использованием ограниченного числа стандартизированных технологических решений, ориентированных на каждую из основных языковых групп. Регионы и страны с более высоким языковым разнообразием требуют, как правило, более сложных подготовительных работ для налаживания Интернет-доступа, которые могут включать адаптацию (кастомизацию) ресурсов для каждого из большого числа миноритарных языков. Исходя из этого, можно сказать, что Интернет изначально «предпочитал» более многочисленные языки. Однако даже крупные языковые группы часто не имеют устойчивых технических стандартов. Так, люди, говорящие на хинди, исчисляются сотнями миллионов, но при этом, по данным одного из исследователей Университета Южной Калифорнии, почти каждый веб-сайт на хинди имеет собственный набор шрифтов, которые несовместимы со всеми остальными наборами шрифтов для этого языка. Люди, желающие прочитать матери-

алы на хинди, представленные на этих веб-сайтах, должны устанавливать шрифты, соответствующие требованиям каждого отдельного сайта. Поиск по этим сайтам проводить очень сложно, т. к. слова, отображенные разным образом, могут быть некорректно распознаны поисковыми машинами (Information Sciences Institute, 2003). Иными словами, не все крупные языковые группы находятся в Интернете в одинаково выгодном положении. Такие регионы как Африка, Океания и Юго-Восточная Азия сталкиваются с еще более серьезными проблемами по причине наличия большого числа языков, не представленных до настоящего времени в Интернете. Как следствие, до того, как включить этих членов языковых групп в число пользователей Интернета, надо проделать большую работу по техническому развитию этих стран.

При рассмотрении влияния Интернета важно не упустить из виду эволюционный взгляд на языковое разнообразие. И хотя Интернет вполне может оказывать долговременное влияние на языковое разнообразие, неясно, каким и насколько большим оно может быть в исторической перспективе. Расширяя сферу действия отдельных языков, Интернет потенциально усиливает их, но, поскольку он проделывает то же самое с большими языками, способствуя одновременно языковым контактам внутри сети, он потенциально ослабляет миноритарные языки. Это влияние может быть намного меньше, чем влияние других общественных факторов, таких как развитие сельского хозяйства, урбанизация населения, геополитические события и пр., которые вполне могут быть вне пределов влияния человека и коллективных организаций, например, ООН, чтобы такое воздействие можно было предотвратить. В то же время, мир является свидетелем настоящего сокращения языкового разнообразия, и под угрозой находится сохранение исторической и культурной традиции сотен сообществ во всем мире. Важно, чтобы эти проблемы были поняты и учтены при формировании любой политики, в задачи которой входят вопросы языкового разнообразия в Интернете.

Источники возникающих сдвигов

Возникающие сдвиги – это предпочтения, появляющиеся при применении Интернет-технологий. По отношению к языковому разнообразию в Интернете возникающие сдвиги появляются в результате работы пользователей информационных технологий, когда знание языков

пользователями начинает соответствовать их возможностям использования предоставленных технологий или информации. Такие сдвиги имеют два основных проявления: во-первых, в распространении языков в Интернете, во-вторых, в экономическом контроле над рынками телекоммуникационных и информационных технологий. В данном разделе мы рассмотрим источники этого вида сдвигов. Представленные здесь данные говорят о существенном предпочтении, отдаваемом английскому языку, и проявления этого сдвига мы наблюдаем в настоящее время.

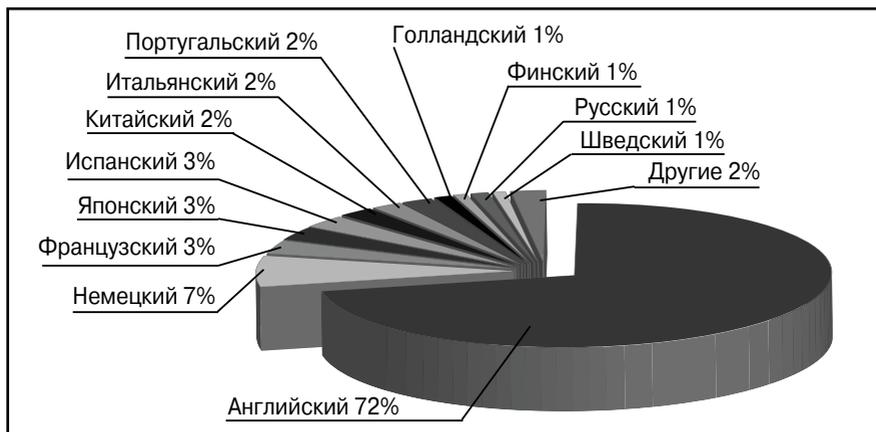
Языковое разнообразие информационных ресурсов Интернета

Не так много исследований посвящено крупномасштабному количественному анализу языков, используемых в Интернете. Как правило, такие исследования фокусируются на Всемирной сети, исключая другие коммуникационные модели, такие как электронная почта и чат, поскольку Сеть легче поддается наблюдению и анализу, чем другие формы Интернет-коммуникации. Два заслуживающих внимания исследования в этой области показали интересные результаты: серия исследований Лавуа, О'Нейлла и др. (Lavoie, O'Neill) из OCLC и исследование Нанберга (Nunberg, 1998) из PARC по неанглоязычным веб-сайтам.

Исследования OCLC (Lavoie and O'Neill, 1999; O'Neill, Lavoie and Bennett, 2003) используют произвольную выборку веб-сайтов в Интернете. Выборка была создана путем генерации произвольных IP адресов, а затем были предприняты попытки выйти на веб-сайты по каждому из этих адресов. Если веб-сервер отвечал, ученые загружали его домашнюю страницу и вводили ее в автоматизированную систему классификации языков (O'Neill, McClain and Lavoie, 1997). Этот метод хорош тем, что исследуется произвольная выборка сайтов. Все другие методы выборки косвенно или напрямую зависят от поисковых машин или «веб-спайдеров» – программ, осуществляющих поиск новых веб-страниц путем следования по ссылкам в известном наборе веб-страниц. Спайдеры создают так называемую выборку по методу «снежного кома», которая определяется близостью к произвольной точке отсчета. Поисковые машины зависят от индексов, создаваемых спайдерами, и такие выборки также будут иметь сдвиги. Если мы хотим получить надежную оценку доминирования различных языков в Интернете, выборку с наличием сдвигов следует избегать.

Начальное исследование тенденций в использовании разных языков проводилось в два разных этапа с интервалом в один год. Более позднее исследование, проведенное в 2002 г., должно было подтвердить полученные наблюдения. Исследование 1998–1999 гг. исходило из предположения, что имела место некоторая международная экспансия Сети и что использование разных языков было тесно связано с доменом, в котором был создан каждый веб-сайт. Выборка, сделанная в 1999 г. и включавшая 2229 произвольно полученных веб-сайтов, содержала 29 идентифицируемых языков, распределение которых представлено на рис. 3. Как и ожидалось, английский язык явно доминировал, и именно на нем было создано 72 % всех исследованных веб-сайтов. Индекс разнообразия для данной выборки веб-страниц составил 2,47, что меньше, чем для типичной страны Юго-Восточной Азии, и больше, чем для типичной страны южной части Центральной Азии. Он также в сотни раз меньше глобального языкового разнообразия. Следовательно, несмотря на то, что языковое разнообразие во Всемирной сети не сильно отличается от ситуации в большинстве многоязычных стран, оно является плохим отражением языкового разнообразия в мире в целом.

Рис. 3. Соотношение языков в Сети по произвольной выборке веб-страниц



Источник: O'Neill, Lavoir and Bennett, 2003.

Второй этап исследования, проведенный в 2002 г., показывает достаточно постоянную долю английского в Сети в сравнении с пре-

дыдущим исследованием, хотя небольшие отличия появляются среди других языков (O'Neill, Lavoie and Bennett, 2003). Индекс разнообразия в 2002 г. оказался равен 2,44, т. е. немного изменился по сравнению с предыдущим исследованием. Частично это могло быть связано с методологией исследования. 29 языков, которые идентифицируются в выборке веб-страниц, представляют собой предел той программы идентификации языков, которую они используют (<http://www-rali.umontreal.ca/SILC/SILC.en.cgi>), и этот метод не позволяет обнаружить новые языки, появляющиеся в Сети. Даже если бы программа идентификации могла охватить большее количество языков, их процент был бы незначителен и поэтому существенно не изменил бы полученный индекс разнообразия языков во Всемирной сети.

Исследование, проведенное OCLC в 1999 г., также определило долю многоязычных веб-страниц каждого домена и языковые пары, используемые на каждом веб-сайте. Если на веб-сайте использовалось более одного языка, английский всегда был одним из них: 100 % из 156 идентифицированных многоязычных сайтов содержали страницы на английском языке. Страницы на французском, немецком, итальянском и испанском языках были выявлены на 30 % многоязычных сайтов, а другие языки составляли еще меньшую долю. Более того, 87 % многоязычных веб-сайтов происходили из доменов вне крупнейших англоязычных стран (Австралии, Канады, Великобритании и США). В рамках каждого из этих доменов уровни многоязычия варьировались от 6 из 13 (42 %) на российских сайтах и до 16 из 1103 (1,5 %) на сайтах США. Следовательно, Всемирная сеть имеет сильную тенденцию к моноязычию, и многие проявления многоязычия – просто «дань вежливости» доминирующему английскому. Это открытие прямо противоречит распространенному мнению, что Сеть, так или иначе, стимулирует разнообразие.

Тенденции, выявленные исследованием OCLC, нашли подтверждение в работе Нанберга (Nunberg, 1998), который использовал другую методику. В его работе была проведена проверка 2,5 млн веб-страниц, отобранных в 1997 г. компанией Alexa, предоставляющей Интернет-услуги. Проверка проводилась с использованием автоматического идентификатора языков, созданного Гейнрихом Шютце (Heinrich Schuetze) – коллегой Нанберга. Несмотря на то, что полученная выборка относится к категории «снежного кома», она в тысячу раз больше вы-

борки OCLC. Основным результатом, полученным Нанбергом, стало следующее: страны с низким уровнем использования Интернета применяли на своих веб-сайтах, в основном, английский язык, а страны с более высоким уровнем использования Интернета, применяли большее количество других языков. Латинская Америка продемонстрировала контрастные показатели по причине очень низкого уровня использования Интернета в 1997 г. и колоссального числа неанглоязычных сайтов. Поэтому степень использования английского как второго языка в неанглоязычных странах может повлиять на языковое многообразие на веб-сайтах этих стран.

В стороне от перечисленных исследований стоят еще несколько попыток изучить распределение языков на основе статистических данных, полученных поисковыми машинами. По разным причинам эти попытки не дали достаточно информации для изучения. Например, FUNREDES – неправительственная организация, занимающаяся вопросами внедрения информационных и коммуникационных технологий в странах Латинской Америки, с 1995 г. проводила серию исследований для оценки распространения языков и национального влияния на Интернет (Pimienta and Lamey, 2001; Pimienta et al., 1995–2003). В ходе этих исследований был осуществлен подсчет веб-страниц, индексируемых наиболее распространенными поисковыми машинами и содержащих определенные слова из разных языков и национальных групп. Интересно, что в результате был получен значительно меньший процент англоязычных страниц (52 % в 2001 г. и 45 % в 2003 г.), чем результат, приведенный в исследованиях Лавуа, О’Нейлла и Нанберга.

Однако подсчет страниц, отобранных поисковыми машинами, представляет собой ненадежную методику определения репрезентативности языков в Сети. Помимо необъективных выборок, в результате которых определенные страницы попадают в поле зрения поисковых машин, существуют и другие схемы, искажающие результаты. Как правило, поисковые машины используют набор методов индексирования индивидуальной разработки, закрытый для проверки, а это может приводить к необъективности при подсчете страниц, результаты которого вы не можете исправить или просто проверить. Ключевое слово не обязательно должно присутствовать на странице, чтобы та попала в подсчет, а страницы, содержащие данное слово, могут выпасть из подсчета. Кроме того, этот метод предполагает, что

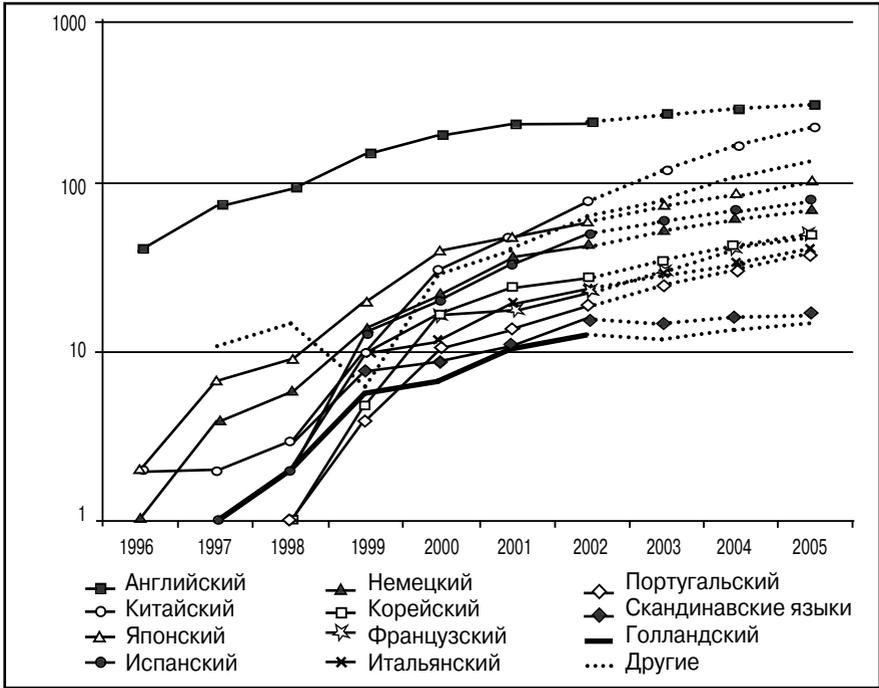
частота встречаемости слов соответствующих «культурно-нейтральных» концептов одинакова в разных языках. Однако культурный нейтралитет недостижим. Многие слова, чью частоту учитывают, представляют культурно-связанные концепты, например, «сыр» (cheese): американская англоязычная культура и культура континентальной Франции придают совершенно разное значение словам, обозначающим этот продукт – cheese и fromage. Данный факт найдет отражение в частоте употребления соответствующих терминов. Более того, если в результате мы получаем число страниц, а не слов, то результат, учитывающий разные формы слова в языке, может включать двуязычные или многоязычные страницы, которые подсчитываются многократно.

Языковое разнообразие пользователей Интернета

Наиболее эффективная попытка оценить языковое разнообразие пользователей Интернета была сделана переводческой компанией Global Reach. Оценки, которые она составляла ежегодно с 1996 по 2002 гг., широко цитируются в качестве иллюстрации мнения об Интернете как средстве развития языкового разнообразия²². Эти оценки базируются на данных Международного союза электросвязи (МСЭ) о количестве пользователей в каждой стране: пользователь определяется как человек, который пользовался Интернетом последние 3 месяца. Группы пользователей делятся по языкам, данные по которым взяты из Ethnologue и сверены с данными ООН по численности населения – т. е. практически так же, как поступали мы, подсчитывая языковое разнообразие (см. выше). В ряде случаев Global Reach дополняла эти источники маркетинговыми данными, полученными от компании Nielsen Net Ratings и некоторых других. В этом исследовании отсутствуют актуальные данные по пользователям Интернета, поэтому данные Global Reach не представляют языки, на которых говорят сегодня пользователи Интернета. Именно потому, что эти цифры часто цитируются как свидетельство языкового разнообразия пользователей Интернета, стоит посмотреть на них внимательнее.

22 Эти данные доступны на <http://global-reach.biz/globstats/evol.html>

Рис. 4. Количество пользователей Интернета, говорящих на различных языках (ось Y дана в логарифмическом масштабе)



Источник: Global Reach.

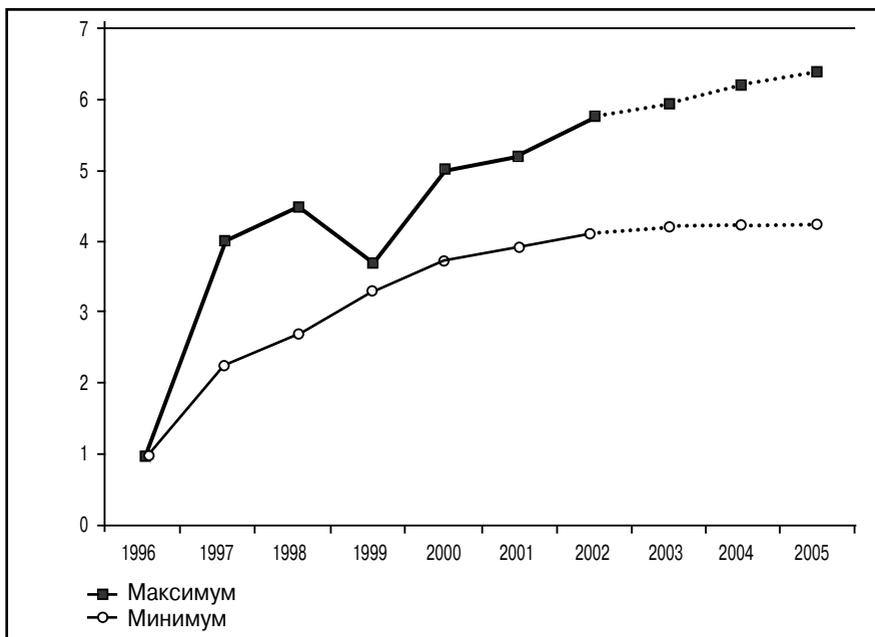
На Рис. 4 представлены данные Global Reach о распределении пользователей Интернета по языкам, на которых они говорят. Период с 2003 по 2005 гг. показан пунктирной линией, поскольку это были прогнозируемые оценки. Установленные исследованием языки пользователей совпадают с языками, определенными OCLC. Как и ожидалось, в 2001 г. количество пользователей, работающих и общающихся в Интернете на английском языке (230 млн человек), в три раза превосходило количество пользователей, читающих и пишущих на китайском языке (60 млн пользователей)²³. Рис. 4 показывает, что все

23 Эти оценки рассматривают все варианты китайского как один язык, несмотря на то, что лингвисты считают китайский язык семьей из 9 разных языков (часто называемых «диалектами» среди нелингвистов).

группы пользователей переживают экспоненциальный рост, за исключением пользователей с английским и японским языками, рост численности которых замедляется. По полученным данным, 50 % от общего количества возможных пользователей Интернета, говорящих на этих двух языках, уже являются реальными пользователями Интернета.

Из оценочных данных, представленных Global Reach, можно вычислить индекс языкового разнообразия для общего числа пользователей Интернета; эти данные представлены на Рис. 5. Поскольку состав языков в группе «Другие» не расшифровывается в данных Global Reach, мы подсчитали минимальные и максимальные значения для индекса, приняв за «другие» один язык (минимальное разнообразие) или равное распределение по 6000 языков (максимальное разнообразие). Удивительно, что, несмотря на значительное увеличение индекса с 1996 по 1999 г., языковое разнообразие, похоже, стабилизируется после 2000 г., хотя и наблюдается экспоненциальный рост многих языков. Кроме того, прогнозы на 2003–2005 г. тоже показывают эту тенденцию выравнивания; прогнозируемый рост числа людей, говорящих на китайском языке, в силу его многочисленности, практически сдерживает рост разнообразия. Конечный результат – это индекс языкового разнообразия, находящийся где-то между типичной африканской страной и суммарного индекса Северной Америки и региональных индексов стран Европы. Ничего удивительного в этом нет, если учесть, что Интернет-хостинг по-прежнему сосредоточен в Северной Америке и Европе. Тем не менее, языковое разнообразие Интернета нигде не превышает индекса какого-то одного региона или мира в целом. С учетом сказанного делаем вывод, что, наперекор популярному убеждению, Интернет в этом смысле не может считаться олицетворением языкового разнообразия.

Итак, глобальный характер и способность соединять огромное количество людей не сделали Интернет лингвистически разнообразным. Чтобы гарантировать представление в сети языков, на которых говорят пользователи Интернета, надо решать и другие вопросы, что, как мы уже видели выше, очень зависит от конкретных сообществ, которые соединяет Интернет.

Рис. 5. Оценка разнообразия пользователей Интернета

Источник: Global Reach

Интернет и практика многоязычия

Доступ к Интернету является обязательным условием использования информации, представленной в Сети. До настоящего момента мы старались понять, что может означать этот доступ в глобальном масштабе. Однако наши усилия не будут успешными, если люди, говорящие на многочисленных языках мира, просто не выберут один из нескольких доминирующих языков. Что определяет языковой выбор пользователей Интернета?

Языки – это не просто средства передачи информации, а сложные системы символов, несущие богатые и тонкие оттенки смысла. Социолингвистические исследования многоязычия говорят о тонкой и неспокойной ситуации существования контактирующих языков, а недавнее исследование многоязычия в Интернете показывает, что такая же ситуация характерна и для Интернета. Более того, глубокий интерес к

Интернету во всем мире основан на экономических преимуществах, которые он предлагает. А способствует ли Интернет развитию языков?

Пользуясь общими терминами, довольно сложно идентифицировать, какие языки и как используются в Сети. Требуется рассмотреть целый комплекс вопросов, начиная с индивидуализации языковых сообществ и заканчивая дифференцированным доступом к Интернету, разными системами письма и компьютерными кодировками, разными коммуникационными режимами. Большинство работ, посвященных возможному влиянию Интернета на язык и культуру, изучают конкретные языковые группы, использующие Интернет в своем специфическом контексте, а не занимаются исследованиями с макросоциальных позиций. Такие конкретные исследования показывают, что языковые контакты в Интернете способствуют развитию больших языков, равно как и контакты вне Сети. Так, Райт (Wright, 2004) и Холмс (Holmes, 2004) приводят данные исследования, посвященного языковому поведению в Сети студентов колледжей, проживающих в восьми разных странах. Результаты показывают, что степень использования родных языков в Сети сильно изменяется в зависимости от исследуемого контекста. В то же время, ни одна из исследованных групп студентов не показала, что использует в Сети свой полный языковой репертуар. Менее распространенные языки вообще не используются в Интернете. Как следствие, вопросы, связанные с этой темой, сложны и туманны.

В своем первом исследовании Паолилло (Paolillo, 1996) открыл, что дискуссионные группы в Usenet, аудитория которой, в основном, говорит на пенджаби, предпочли английский своему родному языку. Такое поведение частично предсказуемо в аудитории, состоящей, в основном, из эмигрантов и людей, получивших образование на английском языке, но наблюдаемые тенденции говорят о том, что пенджаби используется в Сети практически исключительно в ритуальных целях или при общении по национальным вопросам, и служит больше как знак принадлежности, чем как средство информативной коммуникации. В одной из своих следующих работ Паолилло (2006) провел сравнение интерактивности и языковой однородности жителей Южной Азии в ходе их общения в чате и дискуссионных группах в Интернете и обнаружил, что и в том, и другом случае предпочтение отдавалось миноритарному языку (хинди или пенджаби, в зависимости от форума). Эти тенденции получили подтверждение у Пила (Peel, 2004), который показал, что пред-

почтительным языком интерактивных чатов в ОАЭ является арабский язык, а электронные сообщения составляются на английском языке. В другой работе Паолилло (2001) писал, что участники чата, живущие в центре, чаще используют миноритарные языки, чем люди с периферии. Поскольку система чатов позволяет людям включаться в их работу и выходить из нее по собственному желанию, участники с периферии и их языковые предпочтения оказываются доминирующими. Как мы видим, технологические и социальные аспекты Интернет-коммуникации взаимодействуют сложным образом, но все же предпочтение отдается мажоритарным, а не миноритарным языкам. С помощью технических средств можно сгладить последствия языкового доминирования, но неизвестно, насколько это окажется эффективным.

Исследованиями греческого языка в Интернет-коммуникации занимались Кутсогианнис и Митсакопулу (Koutsogiannis and Mitsakopoulou, 2004), Георгакопулу (Georgakopoulou, 2004) и Андросополус (Androtsopolous, 1998). Тематика их исследований охватывает многие вышеперечисленные вопросы. Греческий алфавит, так же как шрифт гурмухи в пенджаби, нелегко использовать в Интернете, а латинизированная форма греческого, адаптированная для внесетевого общения и известная как «Greeklish» (текст на греческом языке, написанный на латинице), встречает поддержку у эмигрантов в многоязычных контекстах, отдающих предпочтение английскому (Georgakopoulou, 2004) или немецкому (Androtsopolous, 1998) языкам. Это, в свою очередь, разрушает греческую норму диглоссии (Ferguson, 1959), при которой говорящие используют отчетливую разговорную форму в неофициальной речи и классический язык в письменной речи. В прошлом правительство Греции прикладывало немало усилий для поддержания грамотности на катаревуса – классическом официальном письменном языке; разрушение греческой диглоссии в Интернете подрывает эти усилия. В другом контексте диглоссии – арабском – Варшауер и др. авторы (Warschauer et al., 2002) отмечали, что разговорный египетский, арабский и английский языки «покушаются» на традиционные функции классического арабского. Такое «покушение» имеет тенденцию дестабилизировать ситуацию с диглоссией, приводя, в конце концов, к предпочтению внешнего, доминирующего языка. В результате, когда в Интернете происходит разрушение лингвистических норм, универсальный доступ к Интернету может иметь разрушительные последствия для языкового разнообразия.

Влияние английского языка одновременно широко и изощренно. Резко контрастные ситуации с электронной почтой в Швейцарии (Durham, 2004) и использованием Интернета в Танзании (Mafu, 2004) показывают, что люди, говорящие на двух языках, отдают предпочтение английскому, а не более близким им местным языкам. Если в употреблении английского языка среди элиты Танзании есть колониальный подтекст, то его нет и не может быть в Швейцарии. Объяснение этому феномену мы находим только в международном статусе английского языка (Crystal, 2003; Phillipson, 1992, 2003). Другим примером влияния английского в Интернете является распространение ряда функций устных языков на письменные языки через SMS, IM (instant messages) и Интернет-чат на шведском языке (Hard af Segerstad, 2002). Точно так же Торрес (Torres, 1991, 2001) отмечал прагматические функции эмоциональных иконок (смайликов) в чате на каталонском языке. Эти формы произошли из англоязычных контекстов и поэтому указывают на контактное влияние английского языка на каталонский через Интернет.

Эти и другие исследования освещают богатство и сложность факторов, влияющих на использование миноритарных языков многоязычными пользователями Интернета. Тема, постоянно появляющаяся во многих исследованиях, касается хрупкости использования недоминирующих языков в коммуникационных контекстах Интернета.

Организации и интересы, управляющие Интернетом

Наперекор популярным убеждениям, Интернет – не открытая и не демократическая (или анархическая) организация. Скорее, это организация со сложной системой мощных интересов, многие из которых сильно централизованы. Эти мощные интересы зачастую не принимают в расчет действия индивидуальных пользователей и, тем самым, создается впечатление, что Интернет свободен от ограничений со стороны гражданских, государственных или корпоративных интересов. Тем не менее, каждый уровень интересов – это возможность определить по языковым сдвигам, какие языки используются в Интернете.

Регулированием Интернета занимаются несколько основных действующих лиц. Во-первых, это телекоммуникационные монополии и олигополии разных регионов мира. Эти компании осуществляют под-

держку инфраструктуры, которая позволяет индивидуальным пользователям подключаться к Интернету, а Интернет-сайтам – соединяться друг с другом. Во-вторых, это компании, производящие компьютеры и программное обеспечение. В их числе Intel, IBM, Hewlett-Packard, Cisco Systems, Sun Microsystems, Microsoft, Adobe. Эти компании создают и продают технические и программные средства, составляющие инфраструктуру Интернета. Кроме того, это управляющие органы Интернета, такие как Корпорация Интернета для специализированных адресов и номеров (ICANN), и Информационный центр сети Интернет (NIC), Американское бюро регистрации адресов (ARIN), Координационный Центр распределения ресурсов сети Интернет в Европейском регионе (RIPE) и Азиатско-тихоокеанский информационный центр сети Интернет (APNIC), которые принимают решения по вопросам подключения к Интернету (см. глоссарий). Национальные правительства также играют определенную роль как в руководстве Интернет-ресурсами на государственном уровне, так и во внедрении других форм информационной политики. И, наконец, есть организации и консорциумы, такие как Консорциум Всемирной сети (W3C), консорциум Юникод (Unicode Consortium) и Международная организация по стандартизации (ISO), которые разрабатывают стандарты для применения технологий Интернета.

Телефонные сети с самого начала были очень важны для Интернета. Когда Интернет-хост соединяется с другим хостом, модемы, выделенные линии, цифровые абонентские линии, оптоволоконные опорные магистрали и геосинхронные спутники – все могут быть задействованы на некотором этапе цифровой коммуникации, физически передавая данные по телефонной сети. За последнее время для передачи Интернет-трафика были адаптированы другие формы телекоммуникационных сетей, такие как телевизионные кабельные сети. Исторически и сегодня экономический контроль над этими ресурсами находится в руках крупных компаний – частных или государственных монополий. Эти концерны наиболее развиты в США. Так, MCI через свое дочернее предприятие UUNET управляет сетью, на которую приходится большая часть международного интернет-трафика (см. Mapnet, <http://www.caida.org/tools/visualization/mapnet>). Оптоволоконная опорная магистраль, введенная MCI несколько лет назад, является главной в этой сети. В то время как компании, аналогичные MCI, практически не интересуются языками, на которых работают Интернет-

пользователи их линий, центральное место, занимаемое США в распределении трафика данных, гарантирует, что административные задачи высокого уровня относительно трафика на опорных магистралях будут решаться на английском языке. Как следствие, региональные сети, связанные с этими центральными сетями, просто обязаны привлекать людей, хорошо владеющих английским языком.

Если на первый взгляд это не представляет собой проблему, поскольку специалисты в области компьютеров во всем мире хорошо владеют английским языком, обе эти тенденции подпитывают и усиливают друг друга. Если руководство региональных сетей не может общаться со своими провайдерами на предпочитаемом ими языке, тогда английский по умолчанию остается доминирующим языком сетевой администрации. Телекоммуникационные компании, собирающие урожай огромных прибылей благодаря спросу на коммуникационные и технологические услуги, несут особую ответственность за языковое разнообразие на рынках, которые они обслуживают.

Компании, производящие компьютерное оборудование и программное обеспечение, оказывают аналогичное влияние на языковой состав Интернета, создавая компьютеры с клавиатурами, мониторами и операционными системами, настроенными на конкретные языки. Эти изделия производятся по низкой цене за счет масштаба производства, и в результате самые крупные рынки мира оказываются заполненными стандартизированной продукцией. Компьютерные технологии с их офшорным производством чипов, разработкой программ за счет аутсорсинга (включая даже менеджмент) и товарными рынками представляют собой один из первых глобальных секторов промышленности. По этой причине и в силу лидирующего положения компаний США в области разработки новых систем и стандартов, компьютерные системы, прокладывающие путь в такие разнообразные в языковом отношении регионы, как Африка, в подавляющем большинстве настроены на использование английского или европейских языков, и почти или совсем не приспособлены к настройке на локальные языки. Это выражается в еще одной форме возникающего сдвига в Интернете в направлении европейских языков, уводящего все дальше от языков тех стран, которые менее развиты в экономическом отношении. Как и телекоммуникационные компании, фирмы, производящие технические средства и программы, несут особую ответственность за языковое разнообразие в тех странах, рынки которых они обслуживают.

Таким образом, действия компьютерных компаний, зажатых конкуренцией за доминирование на рынке, оказывают отрицательное влияние на климат многоязычия в области вычислительной техники и онлайн-языкового разнообразия. Чтобы стимулировать многоязычие в вычислительной технике, нужны меры, предусматривающие приоритет международных интересов над конкурентными целями частных компаний. Некоторые из этих тенденций подкрепляются действиями международных организаций и консорциумов, таких как ISO, Консорциум Юникод, W3C, которые в состоянии предвидеть различные аспекты развития Интернет-технологий. Многие крупнейшие компьютерные компании, включая Apple и Microsoft, работают через эти организации. И неважно, что кто-то из специалистов по технологиям жалуется, что данные организации тормозят инновации, ведь их международный статус помогает принимать во внимание интересы различных национальных и языковых групп. С другой стороны, эти организации по стандартизации не имеют реальных механизмов принудительного применения принятых ими решений. Как следствие, ряд интернет-технологий имеет стандарты, редко соблюдаемые на практике. В их числе использование HTML для веб-страниц и язык программирования ECMAScript для интерактивности веб-браузеров. Несовместимость, возникающая в результате несоблюдения стандартов, разрушительна для развития многоязычия в вычислительной технике. Если мы хотим, чтобы эти организации стимулировали и защищали языковое разнообразие, то нам следует усилить их правоприменительные возможности.

Есть еще одно действующее лицо, управляющее Интернетом и оказывающее большое влияние на его языковое разнообразие, – это ICANN, корпорация, осуществляющая по контракту с Министерством торговли США администрирование протокола, известного под названием «Система доменных имен» (Domain Name System, DNS). DNS осуществляет функцию ассоциирования уникальных мнемонических имен со всеми хостами в Интернете – функцию, которая, в основе своей, является лингвистической. К сожалению, DNS с трудом совмещается с каким-то языком, если только он не является американским английским, и более того, с трудом уживается с функцией присвоения названий на любом естественном языке. DNS глубоко интегрирована в Интернет, поскольку большинство других прикладных протоколов зависят от ее способности находить Интернет-хосты. Это также единственный действительно

контролируемый организацией, а не просто кодифицируемый, протокол. ICANN контролирует DNS, в основном, путем делегирования, но его административная структура, система контрактов с правительством США и другими сторонами, методы работы – все было направлено на ограничение многоязычия в названиях Интернет-хостов. В результате этого DNS не в состоянии выполнить свою изначальную задачу предоставления мнемонических схем для Интернет-хостов. Для изменения этой ситуации требуется изменить ICANN, DNS и политику, проводимую администрацией доменных имен.

Пользователи Интернета воспринимают имена Интернет-хостов почти как обычные имена, а на самом деле, они сильно разнятся. DNS требует, чтобы имена Интернет-хостов были уникальными во всем мире, в то время как в естественном языке метафора, система условных обозначений и аббревиатуры исключают возможность уникальности какого-то конкретного имени. После присвоения домена «acl.org» Ассоциации библиотекарей-христиан (Association of Christian Librarians), это имя становится недоступным для Ассоциации по вычислительной лингвистике (Association for Computational Linguistics) или для любой другой организации в мире, которая хотела бы иметь эту же аббревиатуру в качестве своего названия.

Для обеспечения уникальности в условиях ограниченной гибкости DNS использует иерархически-структурированные имена: индивидуальные имена хостов состоят из цепочек имен, ранжированных от специфичного к более общему. Верхний уровень иерархии занимает последнее поле имени; это будет родовой домен верхнего уровня или код страны (gTLD или ccTLD), который функционирует как общий классификатор. Однако часто остается непонятным, какой классификатор нужно использовать для конкретной цели. TLD, на основании соглашений с ICANN, присваиваются с учетом различных функций: .com предназначен для коммерческих сайтов, .net – для сетей, .org – для неприбыльных и некоммерческих организаций, а коды стран должны присваиваться самими странами с учетом их собственных целей. Доменные имена в gTLD более привлекательны, поскольку, как правило, короче и легче для запоминания. Однако, поскольку количество gTLD незначительно в сравнении с сотнями миллионов хостов, конфликты при присвоении доменных имен неизбежны.

Отношение ICANN к таким конфликтам и к их разрешению строится на поддержке владельцев официально зарегистрированных товарных знаков. В иных случаях первая сторона, регистрирующая свое доменное имя, сохраняет его до тех пор, пока регистрация остается в силе. Это не помогает регистрантам, не имеющим товарного знака или происходящим из маленькой области, или носителям миноритарного языка. Это особенно мешает иностранным заявителям, если именем их национальной идентификации являются омографы ранее зарегистрированного доменного имени. Чтобы изменить имя зарегистрированного домена, требуется провести дорогостоящие переговоры и/или совершить необходимые юридические действия. Регистрация сотен миллионов хостов на английском языке привела к явному сдвигу против интересов неанглийских имен хостов, поскольку тысячи выбранных имен хостов на других языках будут омографами ранее зарегистрированных хостов в gTLD. Как следствие, в DNS присвоение товарного знака как юридически обязательное в США торговое действие получает преимущество над прозрачным многоязычным наименованием, относящимся к сфере языка и международной коммуникации. Такой ассиметричный порядок приоритетов не изменится, пока DNS не будет находиться под управлением полностью международной, а не частной организации, имеющей договорные отношения с правительством США или любой другой страны.

Оригинальная разработка DNS отличается явным техническим сдвигом в направлении предпочтения английского языка, потому что может использовать только 7-битный US-ASCII-код. Поэтому даже европейские языки, такие как французский, испанский и немецкий, использующие диакритику, не отображаемую в US-ASCII, оказываются в затруднении, когда речь заходит о выборе подходящих имен для Интернет-хостов. Ряд организаций, в числе которых Консорциум многоязычных Интернет-имен (MINC), New.net and RealNames долго боролись, стараясь убедить ICANN разработать альтернативы современной системе DNS, обладающие более совершенной поддержкой многоязычия. Несмотря на то, что эти группы внесли много конструктивных, заслуживающих внимания предложений, их усилия встретили большое сопротивление со стороны ICANN. Только недавно ICANN приняла вариант Юникода, известный как «пуникод» (punycode), допускающий существование многоязычных доменных имен, но внедрение этого варианта

проходило неприемлемо медленно, и политически было обречено на провал.

Вопрос доменных имен, в принципе, достаточно символичен. Тем не менее, символизм – это сила, и нетерпимость ICANN к многоязычным доменным именам привела к тому, что мир понял, что этой организации нет дела до интернационализма или языкового разнообразия. Несмотря на то, что в последнее время ICANN была сильно реформирована и сейчас претендует на более интернациональное звучание, она потеряла общественное доверие по вопросу о многоязычных доменных именах, и не ясно, приведут ли эти изменения к справедливой, функциональной и международной системе доменных имен, и можно ли восстановить утраченное доверие.

Роль организаций ARIN, RIPE и APNIC (равно как и других NIC'ов – сетевых информационных центров) в появлении возникающего языкового сдвига не столь очевидна, как роль ICANN. Эти организации с относительно открытым членством осуществляют руководство физическим взаимодействием региональных и локальных сетей. Одной из их основных функций является поддержка пространства IP-адресов. IP-адреса – это 32-битные номера, используемые для уникальной идентификации хостов. Как и доменные имена, IP-адреса присваиваются в процессе делегирования посредникам, которые могут передавать полномочия дальше. В отличие от доменных имен, каждый присвоенный диапазон соответствует физической ветви Сети, соответствующее оборудование которой управляется одним провайдером. IP-адреса присваиваются в диапазонах, а поскольку адресное пространство, в конечном итоге, лимитировано, каждое такое присвоение имеет цену возможности – такие же номера не могут быть присвоены позднее где-то в другом месте, если только эта часть Сети не прекращает своего существования.

Сетевые информационные центры (NIC) вынуждены заниматься вопросами языкового разнообразия, выполняя возложенные на них функции региональной администрации. Сетевые ресурсы, доступные конкретной стране или языковой группе, зависят от диапазона IP-адресов, предоставленных соответствующему региональному администратору, и их передачи от него другим группам и странам. Плохая передача адресов или малый диапазон, с которого начинается работа, представляют собой два препятствия, которые могут привести к недостатку

адресов для новых хостов. Разгорелась полемика, имеет ли APNIC, региональные обязанности которой охватывают Океанию, Восточную и Юго-Восточную Азию, достаточно места для адресов, чтобы продолжать присвоение IP-диапазонов с необходимой скоростью. APNIC отрицает существование такой проблемы, но призрак надвигающегося кризиса вызывает тревогу. Как ожидается, проблемы, связанные с адресным пространством, могут усугубиться в результате модернизации текущей 4-й версии IP (IPv4) и появления 6-й версии IP (IPv6), которая использует более широкий диапазон адресов; но эта конверсия была отложена на несколько лет по причине технической несовместимости с IPv4.

Тем не менее, присвоение адресного пространства IPv4 очень неэффективно. Большие диапазоны адресного пространства выделяются для специальных целей или остаются полностью неиспользованными; они известны как «bogons» и находятся под строгим контролем для того, чтобы системные администраторы могли осуществлять их мониторинг в целях безопасности (см. <http://www.cymru.com/Bogons/>). Даже когда диапазоны «bogon» были выведены из области использования, произвольная выборка из 1107 IP-адресов возвратила 203 IP-адреса (18 %), которые, очевидно, были отданы под тестирование редко используемого протокола многоадресной передачи. Иными словами, 18 % глобально доступного пространства IP-адресов были заблокированы и стали неиспользуемыми вследствие неэффективности передачи этого пространства. Именно потому, что такая неэффективность может иметь место, и потому, что она может оказывать влияние на адресное пространство, выделяемое региональным администрациям, локальные языковые группы могут не получить доступ к Интернет-ресурсам. Чтобы разные языки имели равный шанс на использование в Сети, администрирование и выделение адресного пространства в Интернете должно осуществляться справедливо.

Национальные правительства могут играть как положительную, так и отрицательную роль при оказании влияния на языковые сдвиги в Интернете. Когда национальные правительства принимают внутри своих стран политику, защищающую и развивающую языковые права своих многоязычных граждан (Skutnabb-Kangas and Phillipson, 1995), исторические языковые сдвиги в этих странах находятся под контролем. Поскольку эта языковая политика переносится на соответствующие области информационной политики, она

способствует развитию языкового разнообразия в Интернете. Однако, как правило, правительства больше озабочены административной стороной деятельности и опасностями сепаратизма, и значительная часть населения мира живет без гарантий обеспечения своих базовых языковых прав. Когда страны контактируют с глобальным Интернетом и требуют приспособить его к своим национальным языкам, они стимулируют появление возникающего сдвига, в ущерб своим собственным этнолингвистическим меньшинствам, принося, в конечном итоге, мало содействия языковому разнообразию в сети. Если национальные языковые группы надеются обеспечить себе нишу в глобальной телекоммуникационной этносфере, они должны признать языковое разнообразие и заниматься им в пределах своих национальных границ. Они могут также стремиться к обучению компьютерной грамотности граждан всех своих языковых групп, поскольку это необходимо для полноценного использования Интернета. Этнолингвистическое понимание телекоммуникационных компаний, компьютерных компаний и руководящих органов Интернета начнет расширяться только тогда, когда критическая масса недостаточно представленных этнолингвистических групп сможет привлечь к себе внимание. Это вряд ли произойдет, пока скрыт истинный размах международного языкового разнообразия.

Возникающие языковые сдвиги – важный аспект, требующий к себе особого внимания при рассмотрении вопросов языкового разнообразия в Интернете. Проблемы, рассмотренные нами в настоящей статье, носят чисто иллюстративный характер и не претендуют на представление исчерпывающего списка возможных возникающих сдвигов. По мере изменения рынка телекоммуникаций, компьютерной техники и программного обеспечения, по мере развития организаций, осуществляющих руководство Интернетом, могут появляться новые языковые сдвиги. Возникающие в разных технических и лингвистических контекстах языковые сдвиги могут быть глубоко локальными, поскольку проявляются особым образом внутри конкретной страны. Именно поэтому общая проблема возникающих языковых сдвигов требует тщательного мониторинга на глобальном, региональном и локальном уровнях.

Источники технических сдвигов

Три области технических сдвигов, имеющих разное отношение к языковому разнообразию, заслуживают внимания в свете усилий по интернационализации в рамках трех рассмотренных выше направлений действий ЮНЕСКО. Во-первых, это вопрос стандартов кодирования, который имеет прямое отношение к действиям по п. 10, стимулируя языковое и культурное разнообразие в Интернете. Кодирование текстов – основное техническое средство достижения языкового разнообразия в этой базовой, основанной на текстах, коммуникационной среде. Во-вторых, это вопрос языков разметки и программирования, используемых для создания и поддержания Интернет-приложений и контента. Эти технические системы напрямую связаны с п. 9 – продвижением компьютерной грамотности. Если компьютерная грамотность в качестве обязательного условия требует грамотности в другом языке, открытость и универсальный доступ гарантировать невозможно. Наконец, это вопросы, связанные с техническими языковыми сдвигами в протоколах Интернет-приложений – пп. 9 и 10. Для содействия доступу развивающихся стран к информационным технологиям основные Интернет-приложения (электронная почта, гипертекстовые браузеры, мгновенный обмен сообщениями в сети и пр.) должны разрешать использование языков этих стран. Без этого барьеры на пути принятия технологий станут непреодолимыми. Рассмотрим подробнее три названные области технических сдвигов.

Кодирование

Кодировки определяют произвольное присвоение чисел символам письменных языков. Два разных вида кодировок могут быть несовместимыми в случае присвоения одного числа двум разным символам и наоборот. Чтобы воспользоваться возможностями, которые предоставляет нам компьютер при работе с текстом (по его отображению, редактированию, сортировке, поиску и быстрой передаче), коммуникация на определенном языке должна быть представлена в некой кодировке. Поэтому большая часть из того, что предлагает нам Интернет в отношении языкового разнообразия, сводится к кодировкам конкретного текста.

Наиболее распространенной кодировкой является ASCII код (American Standard Code for Information Interchange), разработанный в 1950–1960-е гг. под эгидой ANSI – Американского института нацио-

нальных стандартов – для стандартизации технологии работы по теле-тайпу. Этот код состоит из 128 символов и пригоден, в основном, для североамериканского английского. По причине его раннего появления и широкого распространения все коды, появившиеся позже, строились с ориентацией на ASCII, например, ISO-8859-1 или Latin-1 Международной организации по стандартизации. Этот код содержит 256 символов, первые 128 из которых совпадают с символами ASCII. Юникод – попытка предоставить совместимые кодировки для всех языков мира (Unicode Consortium 1991, 1996, 2000, 2003) – принял на вооружение такую же стратегию, сделав первые 256 из 65 536 символов в BMP (Базовой многоязыковой плоскости) такими же, как в ISO-8859-1. Большинство технологий, поддерживающих Интернет, основано на ASCII или его производных. DNS, Usenet news и Internet Relay Chat, разрешают использование только поднабора символов ASCII. Операционные системы, такие как Linux, используют исключительно «текстовые ASCII-файлы» для ряда своих базовых функций. Все эти системы усиливают технический сдвиг в направлении английского языка.

Надежды на интернационализацию инфраструктуры Интернета связаны с окончательным принятием Юникода – попыткой стандартизации, предпринятой Консорциумом Юникод в сотрудничестве в ISO. Членами Консорциума являются крупнейшие поставщики программного обеспечения, представители мировых религий, региональные организации и учреждения образования, правительственные организации разных стран. Стандарт Юникод (в нынешней 4-й версии) представляет более миллиона возможных кодов символов, что позволяет использовать в едином тексте все современные и исторические письменности. 65 000 символов составляют Базовую многоязыковую плоскость, которая должна обеспечить бóльшую часть письменной коммуникации. Однако такое разнообразие стоит денег. В своей базовой форме UTF-32 текст в Юникод занимает в 4 раза больше места в сравнении с аналогичным текстом в ASCII. Многие разработчики программ предполагали, что пользователи не захотят получить такое «наказание» за пользование многоязычным текстом, особенно если работа на компьютере происходит в монологичном контексте²⁴. Юникод предлагает другие виды кодирования с символами переменной длины, которые бо-

24 Так это или нет – важный вопрос, не получивший достаточного рассмотрения в научной литературе.

лее эффективны, но ценовое бремя переносится на нелатинские языки, которые вынуждены занимать больше места. И хотя за последние 10 лет цены на хранение данных сильно упали, позволяя снизить проблемы Юникода, работа с этой кодировкой по-прежнему сильно усложняет задачу разработчикам программ, поскольку большинство приложений требуют взаимодействия с ASCII. Кроме того, большой размер документов в Юникоде требует больших затрат на пересылку, сжатие и распаковку, и эти расходы – достаточное основание для того, чтобы в ряде случаев отказаться от использования Юникода.

Несмотря на крупные изменения в деле интернационализации обработки данных с помощью Юникода, проблемы с использованием многоязычного текста в Интернете еще далеко не решены. По ряду технических, экономических и организационных причин разработка приемлемого технического стандарта идет медленнее, чем развитие самого Интернета. Как следствие, международное использование Интернета поддерживало языки, основанные на латинице (английский язык в особенности), чему способствовало наличие широко распространенного стандарта кодировки, принятого до распространения Интернета. Чтобы в Интернете можно было одинаково использовать все языки мира, Юникод должен получить более широкое распространение. Как и в случае с DNS, может потребоваться модернизация ряда Интернет-протоколов и приведение их в состояние, совместимое с Юникодом.

Языки разметки и программирования

Есть и другой инструмент, благодаря которому в Интернете сохраняются технические сдвиги в пользу английского языка, – это компьютерные «коды», т. е. языки разметки и программирования, которые используются для конфигурации Интернет-контента и услуг. Первым и наиболее очевидным способом создания технических сдвигов является поддержка этими языками многоязычного контента. Языки разметки, такие как Hypertext Markup Language (HTML) и eXtensible Markup Language (XML), должны уметь описывать текст на всех возможных человеческих языках. Консорциум Всемирной сети предпринял для этого ряд шагов, потребовав поддержать Юникод как элемент своих стандартов. Это означает, что там, где поддержки Юникода недостаточно, поддержки HTML и XML тоже недостаточно. Именно это происходит сейчас с большинством языков Западной и Юго-Восточной Азии, а также южной части

Центральной Азии. Так создается общая основа для предпочтения определенных языков. Языки программирования также должны быть совместимы с многоязычным текстом. К сожалению, многие распространенные языки программирования, такие как С, пока еще не обеспечивают стандартной поддержки Юникода²⁵. Растущее число языков, созданных для веб-приложений (в их числе широко используемые Java, JavaScript, Perl, PHP, Python, Ruby), способны на поддержку Юникода, но другие системы, как, например, программное обеспечение для баз данных, непостоянны в своей поддержке Юникода. Организация электронной коммерции на различных языках, а не только на английском, предполагает, что совместимые с Юникодом базы данных получат широкое распространение.

Следующий путь создания сдвига в сторону английского языка – дизайн самих языков разметки и программирования. Языки программирования предлагают самый простой из имеющихся интерфейс для управления компьютером, который является посредником между когнитивными процессами программистов и логическими возможностями самих компьютеров. Число языков программирования огромно – от 2500 до количества, превышающего число человеческих языков. Несмотря на такое очевидное разнообразие, подавляющее большинство этих языков берет свое начало в FORTRAN – самом первом из языков программирования высокого уровня, созданного IBM в 1957 г. (Levenez, 2003). Эти языки широко используют английские ключевые слова для определения важных компонентов программирования, таких как условные операторы (if, then, else, case и др.) и итерационные циклы (while, for, until и др). Несмотря на то, что во многих человеческих языках есть эквиваленты этих слов, они никогда не используются вместо этих английских слов в выполняемой программе. Так, Руби, работающий в соавторстве с японским программистом Юкихио Матсумото и обращающий особое внимание на интернационализацию, также использует английские ключевые слова.

HTML и XML в этом отношении очень похожи. Тэги HTML представляют собой, как правило, мнемонические аббревиатуры английских

25 Веб-сайт «International Components for Unicode» предлагает C-библиотеку с открытым кодом, поддерживающую использование Юникода (<http://oss.software.ibm.com/icu/>).

слов (например, **b** для **bold** – полужирный шрифт, **ul** для **unordered list** – неупорядоченный список, **li** для **list item** – элемент списка и др.). И, хотя XML не является языком разметки, это – синтаксис для определения языков разметки, все языки разметки, базирующиеся на XML, используют английский (например, MathML для математических выражений и XML: FO для форматирования текстовых документов), несмотря на стандарт XML, использующий Юникод. Эта тенденция усугубляется в связи с проектом развития Семантического веба («Semantic Web»), цель которого в том, чтобы привнести во Всемирную сеть идеологию «общеизвестного знания». Предполагается, что крупные базы данных искусственного интеллекта, такие как Cyc (Reed and Lenat, 2002) и WordNet (Fellbaum and Miller, 1998), будут использоваться для разработки новых языков разметки, и эти языки будут способствовать тому, что Интернет-программы смогут находить и обрабатывать информацию для пользователей. Такие базы данных уже подверглись критике с позиций культуры стран Северного полушария за сексистские, патриархальные предпочтения (Adam, 1998). Именно поэтому проекты типа Семантического веба, которые обещают дать жизнь «новому поколению» Интернет-услуг, несут в себе опасность углубления существующих языковых и культурных сдвигов.

Потенциальную опасность языковых сдвигов, заложенную в языках программирования и разметке, следует анализировать параллельно с культурной природой вычислений. Современная система вычислений основывается на многовековой истории математических знаний, и распространение этой системы на современном этапе можно сравнить по характеру и по значению с распространением системы десятичного исчисления. Десятичные числа были изобретены в Северной Индии около VII века н.э. и распространились по всему миру, вытесняя другие системы исчислений. При этом культурная передача десятичных чисел не требовала импортирования словарного состава; многие языки модифицировали свои названия чисел для восприятия нового метода. Компьютер развивает принципы десятичных чисел путем автоматизации работы с ними. Однако в отличие от десятичных чисел распространение компьютеров принесло с собой объемные и сложные английские словари – языки программирования.

Нет сомнения, что компьютер, как физический артефакт, привязывая символы к действиям, играет определенную роль в их взаимосвязи. Точное сочетание символа и действия имеет произвольный характер,

и поэтому адаптировать можно любой язык, но это – довольно сложная задача, выполнение которой далеко не тривиально. Как следствие, один важный вопрос, касающийся языкового разнообразия, не получил должного внимания в научной литературе, а именно: до какой степени различные характеристики языков программирования способствуют тому, что их воспринимают и используют люди, говорящие на разных языках?²⁶ Хорошо известны эффекты переноса у людей, говорящих на одном языке и изучающие другой. Вполне естественно, что языки программирования, как формальные лингвистические системы, могли проявлять существующий в естественных языках перенос, создавая систематические проблемы или ошибки у людей, говорящих на определенных языках. По своим свойствам языки программирования сильно отличаются друг от друга. Возможно ли, что люди, говорящие на каком-то языке, получают более качественные услуги от того языка программирования, который больше соответствует их родному языку? Тогда, может быть, языки программирования могут быть составлены с учетом различных культурных и языковых традиций? Поможет ли такая адаптация людям, говорящим на каком-то определенном языке, освоить ресурсы информационных технологий?

ЮНЕСКО и другие агентства ООН испытывают настоятельную потребность увидеть ответы на эти вопросы, особенно тогда, когда для продвижения языкового разнообразия необходимо решать задачи в области образования. Компьютерное программирование сделало язык живым и сильным, способным видоизменять культуры. Жаль, что сегодня такую поддержку получил, в основном, английский язык. Если цифровая грамотность в области языков программирования требует языковых и/или культурных знаний на английском, тогда люди, говорящие на других языках, должны нести большие расходы на образование и, возможно, на культуру, чтобы заявить о своем праве на информационные ресурсы Интернета.

Коммуникационные режимы

Несмотря на то, что Интернет стал известен большинству людей благодаря Всемирной сети (а многие считают, что это одно и то же), он пред-

26 См. Anis (1997) для ознакомления с предложениями по данному вопросу.

ставляет собой более разнородную среду, предлагающую различные варианты коммуникации. Более того, дизайн Интернета подразумевает возможность создания и недорогого разворачивания новых вариантов. Сегодня мы широко используем электронную почту, Сеть и мгновенный обмен сообщениями в Интернете и не знаем, что мы будем использовать в будущем. Определенные коммуникационные режимы стали очень популярны и периодически инкорпорируют технические формы языковых сдвигов.

Одним из таких коммуникационных режимов является Usenet News, изначально созданная в 1978 г. как сеть трех университетских компьютерных систем (Spencer and Lawrence, 1998). Usenet – это коллекция тысяч «новостных групп», пространство для публичных сообщений с именами, позволяющими отнести их к определенной теме. Сервисное и клиентское программное обеспечение Usenet имеется в свободном доступе, и его администрация относительно открыта. Администраторы Usenet могут в индивидуальном порядке устанавливать объем, скорость и частоту обмена сообщениями с другими серверами, так что они могут быстро оптимизировать пользование сетью в районах с плохой связью. Таким образом, барьеры для доступа в Usenet относительно невысоки. Usenet – чрезвычайно важный международный ресурс. По состоянию на 1999 г. доступ в Usenet имели 205 стран (Smith, 1999).

Как техническая система, Usenet – это микрокосм Интернета. Система имен для новостных групп носит иерархический характер и использует подсистему ASCII, почти так же, как DNS. Она имеет иерархии верхнего уровня и локальные, региональные и национальные иерархии²⁷. Текст сообщений должен быть совместим с ASCII. Тексты на китайском и японском языках используют в Usenet особые кодировки. Как и в остальной части Интернета, имеет место преобладание английского языка в родовых иерархиях верхнего уровня. Так, в иерархии comp. – родовой категории записей, касающихся компьютерных систем, почти нет записей на японском языке, даже на comp.lang.ruby. Только в иерархии fj.comp. можно найти дискуссии по вычислительной технике на японском языке. В ветви иерархии soc.culture. есть место для многоязычного трафика, но, в основном, на европейских языках.

27 Пространство имен в Usenet, как и пространство имен в DNS, было использовано с серьезными нарушениями.

Таким образом, несмотря на низкую стоимость для стран с ограниченными ресурсами, Usenet плохо приспособлена для международного использования и обладает множеством технических сдвигов в направлении английского языка. Некоторые из этих сдвигов приводят к дополнительным возникающим сдвигам.

Второй коммуникационный режим, ставший популярным в начале 1990-х годов, – ретранслируемый Интернет-чат (Internet Relay Chat, IRC), многосторонний синхронный режим коммуникации в реальном времени. Участники этого канала коммуникации общаются между собой в реальном времени почти так же, как если бы они участвовали в телефонной конференции, с той лишь разницей, что разговоры записываются. Соединенные в общую сеть, серверы IRC могут объединять тысячи каналов, и эта сеть, как и EFNet или UnderNet, открывает культурные, региональные или национальные темы и привлекает участников со всего мира (Paolillo, 2001). IRC был создан в Северной Европе, так что некоторые его характеристики, такие как допустимые символы в текстовых сообщениях и имена участников, отличаются от принятых в Usenet. Однако поддержка многоязычных текстов в IRC несколько не лучше, чем в Usenet. Фактически, отображение текстов на американском варианте английского языка на компьютерах в странах Северной Европы обнаруживает очевидные проблемы, в числе которых замена знаками препинания символов диакритических гласных в скандинавских именах и словах.

Итак, несмотря на притязания этих двух систем на международное использование, они обладают недостатками, которые берут свое начало в языковых сдвигах, составляющих часть дизайна этих систем. Понятно, что новые коммуникационные режимы, такие как Instant Messaging, Web-logging, Web-chat и другие, появляются постоянно. Несмотря на то, что часть из них изначально инкорпорирует XML и Юникод, состояние этих стандартов таково, что только небольшая часть населения мира и мировых языков может пользоваться услугами этих технологий. Многие сторонники технического развития возлагают надежды на новые коммуникационные протоколы, такие как voice-over-internet (голос через Интернет) или multi-modal interfaces (мультимодальные интерфейсы). Но даже если они смогут решить какие-то языковые вопросы, останутся другие проблемы, и в их числе поддержка людей с дефектами слуха и зрения. Более того, существующие технические сдвиги усиливают сдвиги, появляющиеся по демографическим, экономическим

и другим причинам. Чтобы минимизировать эти явления в Интернете, новые коммуникационные режимы должны быть внимательно изучены на предмет наличия латентных технических сдвигов до того, как они будут предложены для массового использования.

Многие «технофилы» возлагали надежды на машинный перевод как на решение проблем многоязычной коммуникации в Интернете. Уже сейчас существует большой спрос на переводческие услуги, предлагаемые компаниями типа Systran – поставщиком системы перевода BabelFish, – и в определенных ситуациях, например, в случае перевода с каталонского языка на испанский и обратно, машинный перевод решил серьезные коммуникационные проблемы (Climent et al., 2004). Так смогут ли люди получать доступ в Интернет на своем родном языке, просто используя одну из онлайн-овых систем перевода? Этот вопрос носит слишком оптимистический характер – и причин тому несколько.

Во-первых, наличие системы машинного перевода предполагает, что «мирские» проблемы представления и понимания текста на определенном языке уже решены, в то время как для большинства языков мира это не так. Во-вторых, создание систем машинного перевода слишком трудоемко. При переводе любой пары языков могут возникнуть специфические проблемы, которые решаются только для этой конкретной пары языков. Так что перевод с любого языка на любой другой язык мира – это задача, которая вряд ли будет решена в ближайшем будущем. В-третьих, создание систем машинного перевода требует огромного объема материалов на языках, с которых и на которые будет осуществляться перевод²⁸; на сегодняшний день эти материалы собирают с веб-сайтов на тех языках, на которые будет реализован перевод (Grefenstette, 1999; Resnik, 1999), и поэтому они должны быть составлены носителями этих языков. Такая работа не будет сделана, пока не будет подготовлена адекватная техническая поддержка для каждого конкретного языка. Наконец, машинный перевод никогда не будет так же хорош, как перевод, выполненный человеком (Kau et al., 1993). Пользователи систем машинного перевода должны приспосабливаться к странным обрывкам слов и их порядку, которые просто представляют

28 Для этого нельзя использовать любой текст. Как правило, требуются двуязычные тексты, сопоставленные по каждому предложению. Их создание стоит довольно дорого, и для всех пар языков таких текстов нет.

собой скрытую форму языкового сдвига, собственно, и приведшего к необходимости перевода. Следовательно, не приходится ожидать, что такие технологические подходы, как машинный перевод, сумеют существенным образом сократить проблемы языковых сдвигов в Интернете.

Выводы

Изучение потенциальных источников сдвигов, рассмотренных в данной статье, вскрыло большое количество источников языковых сдвигов в Интернете, как исторических и технологических, так и возникающих. Следовательно, можно ответить утвердительно на вопрос, существуют ли языковые предпочтения в Интернете. Основные последствия сдвигов заключаются в благоприятствовании большим языкам с хорошо развитыми техническими стандартами, самым значительным из которых является английский язык, и это не случайно, т. к. именно английский был языком основателей Интернета и предшествующих проектов. Однако также очевидно, что причины и следствия сдвигов разнообразны, неявны и, во многих случаях, непредсказуемы. Если ЮНЕСКО серьезно надеется решить проблему языковых сдвигов в Интернете, нам надо больше работать над собственным образованием и образованием отвечающих за развитие Интернета основных действующих лиц в двух сферах: проявления языковых сдвигов, и значение и ценность языкового разнообразия.

Глоссарий

Возникающий сдвиг (Emergent bias) – по Фридману и Ниссенбауму (Friedman and Nissenbaum, 1997) – это предпочтение, которое является следствием взаимодействия технических систем в разных социальных контекстах.

Всемирная сеть (Сеть, World-Wide Web, Web) – приложение для обмена форматированными документами, программами и мультимедийным контентом по Интернету. Это также весь набор документов и контента, доступных в Сети. Сеть – самое популярное приложение Интернета, благодаря простоте работы веб-браузеров при обработке запросов на документы и другого контента.

Интернет (Internet) – глобальная компьютерная сеть, возникшая в результате слияния ARPANET и других региональных компьютерных сетей.

Исторический сдвиг (Pre-existing bias) – по Фридману и Ниссенбауму (Friedman and Nissenbaum, 1997) – любые предпочтения, вызванные социальными причинами и возникшие до появления конкретного технологического приложения, в котором они проявились.

Консорциум Юникод (Unicode Consortium) – организация, осуществляющая контроль над развитием Юникода.

Протокол (Protocol) – стандартизированный набор сообщений и правил для обмена сообщениями между компьютерами в Сети. Протоколы – это сложные системы, часто обсуждаемые в терминах т. н. «уровней»: прикладной уровень, канальный уровень и т. п.

Регистратор доменных имен (Domain-name registry) – организация, которая по контракту с ICANN осуществляет руководство некой частью пространства имен DNS. Как правило, регистратор отвечает за ведение одного или более TLD. Примерами регистраторов доменных имен являются Verisign и Educause.

Технический сдвиг (Technical bias) – по Фридману и Ниссенбауму (Friedman and Nissenbaum, 1997) – любые предпочтения, являющиеся составной частью технической системы. Предпочтение в ASCII американского варианта английского языка является типичным примером технического сдвига.

Техно-лингвистическая жизнеспособность (Techno-linguistic vitality) – возможность этно-лингвистического сообщества воспользоваться преимуществами технологий, в особенности информационных технологий, и возможность использовать родной язык при работе с этими технологиями. По аналогии с этно-лингвистической жизнеспособностью.

Хост (Host, Internet host) – любой компьютер, подключенный к Интернету.

Юникод (Unicode) – 64-битный стандарт кодирования символов, находящийся на стадии развития и позволяющий представить знаки практически всех письменных языков. Разработкой Юникод занимается Консорциум Юникод совместно с ISO и W3C, что должно обеспечить совместимость стандартов всех трех организаций.

Язык разметки (Markup Language) – система, внедряющая форматирование или другие коды («разметку») в текстовые документы, позволяющие форматировать текст или интерпретировать его с помощью устройства, понимающего разметку. Примером языка разметки является HTML (язык гипертекстовой разметки); другие языки, такие как SVG (Scalable Vector Graphics – масштабируемая векторная графика), работают по сходной схеме, но выполняют совершенно разные функции. См.: XML.

АСМ (Association for Computing Machinery) – Ассоциация по вычислительной технике – крупнейшая международная профессиональная организация в области вычислительной техники. Ассоциация образует множество специальных групп, которые занимаются техническими, социальными и политическими вопросами, имеющими отношение к компьютерам и компьютерным сетям.

АPNIC (Asia-Pacific Network Information Center) – Сетевой информационный центр, осуществляющий контроль за работой Интернета в Азии и Тихоокеанском регионе. Под его контролем находятся Австралия, Китай, Япония, Корея, Индонезия, Малайзия и все страны островов Тихого океана.

АRIN (American Registry for Internet Numbers) – американское бюро регистрации адресов в Интернете. Это сетевой информационный центр, который осуществляет контроль за работой Интернета в Северной Америке.

ASCII (American Standard Code for Information Interchange) – один из первых 7-битных стандартов кодирования компьютерных текстов, который получил повсеместную поддержку в большинстве компьютерных приложений. Большинство современных текстовых кодов, включая Юникод, совместимы с ASCII, сеть битов которого позволяла кодировать 128 разных символов. Расширенный ASCII – это 8-битный вариант прежнего ASCII, который не имеет стандарта; разные поставщики поддерживают разные версии этого расширенного ASCII, которые друг с другом не совместимы.

BMP (Basic Multilingual Plane) – Базовая многоязыковая плоскость, область значений кодов Юникод, содержащая символы практически для всех современных письменностей и большое число специальных символов.

ccTLD (Country-Code Top-Level Domain) – домены верхнего уровня, которые ассоциируются с определенными странами. ccTLD идентичны ко-

дам стран ISO-3166. В качестве примеров можно привести .uk (Великобритания) и .za (ЮАР).

СМС (Computer-Mediated Communication) – компьютерно-опосредованная коммуникация – любая форма коммуникации «человек-человек», которая происходит при помощи сетевых компьютеров как средства коммуникации.

CNNIC (China Network Information Center) – китайский сетевой информационный центр, осуществляющий контроль за работой Интернета в Китае.

DNS (Domain-Name System) – система доменных имен – техническая система, которая находится под управлением ICANN, что позволяет присваивать мнемонические коды компьютерам, объединенным в Сеть.

Ethnologue – база данных, ведение которой осуществляет SIL International (редактор – Barbara Grimes). База содержит записи, представляющие общую дескриптивную информацию всех известных языковых популяций мира.

GPL (Gnu Public License) – юридическое лицензирующее соглашение, которое используется в ряде программ с открытым кодом и предназначено для защиты авторского права, предоставляя при этом для разработчиков открытый доступ к исходному коду программ.

gTLD (Generic Top-Level Domain) – родовой домен верхнего уровня. Домены верхнего уровня присваиваются «родовым» категориям, без обязательной ссылки на страну. Общие gTLD – это .com (коммерческие), .edu (высшие учебные заведения США), .mil (организации военного ведомства США), .net (сетевые провайдеры), .org (неприбыльные и некоммерческие организации) и т. д.

HTML (Hypertext Markup Language) – гипертекстовый язык разметки, который используется для форматирования страниц во Всемирной сети. Это – простой язык разметки, понимаемый широким спектром веб-браузеров и другим программным обеспечением, со стандартом, который теперь поддерживается W3C.

IANA (Internet Assigned Numbers Authority) – уполномоченная организация по распределению нумерации в Интернете.

ICANN (Internet Corporation for Assigned Names and Numbers) – Корпорация Интернета для специализированных адресов и номеров, контролирующая выдачу доменных имен верхнего уровня и отвечающая за адресное пространство Интернета.

ICT (Information and Communications Technology) – информационно-коммуникационные технологии.

IP (Internet protocol) – Интернет-протокол. См.: TCP/IP.

IPv4 – IP версия 4. Это версия IP, имеющая сегодня самое широкое применение. Она характеризуется 32-битными номерами адресов для каждого хост-компьютера Интернета. Адресное пространство под IPv4 ограничено и, как следствие, Интернет сегодня переживает переход от IPv4 к IPv6.

IPv6 – IP версия 6. Это – версия «следующего поколения» Интернет-протокола с использованием 128-битных адресов. Поддержка IPv6 расширяется в целом ряде сетевых приложений, но широкого применения она пока еще не нашла, поскольку приложения IPv4 не совместимы с хостами IPv6.

IRC (Internet Relay Chat) – протокол приложений для одновременной многосторонней коммуникации в Интернете в режиме реального времени. В программах «чатов», включая и частные программы, есть много заимствований из IRC. Существует множество сетей IRC, которыми пользуются миллионы людей во всем мире, в основном, в досуговых целях.

ISO-8859-1, Latin-1 – 8-битный стандартный текстовый код, поддерживающий большинство европейских языков, использующих латинский алфавит.

NIC (Network Information Center) – сетевой информационный центр, техническая организация, контролирующая работу Интернета на региональном или местном уровне. Существует 3 основных региональных центра: ARIN, RIPE, APNIC – для Северной Америки, Европы и Азиатско-Тихоокеанского региона соответственно.

RIPE (Réseaux IP Européens) – Координационный Центр распределения ресурсов сети Интернет в Европейском регионе, сетевой информационный центр, контролирующий работу Интернета в Европе.

SGML (Standard Generalized Markup Language) – стандартный общий язык разметки, широко использующийся в печатной и издательской сфере. Язык HTML был изначально разработан как приложение SGML.

TCP/IP (Transmission Control Protocol/Internet Protocol) – набор сетевых протоколов, на которых базируется Интернет. TCP и IP – это практически независимые «уровни» сетевых Интернет-протоколов, поскольку отвечают за разные аспекты функционирования сети, но часто используются вместе.

TLD (Top-Level Domain) – домен верхнего уровня. Доменное имя, присваиваемое ICANN регистратору доменных имен, который объединяет взаимосвязанные хосты, как правило, в рамках одной страны или в организационных целях.

Usenet (news) – компьютерная сеть, используемая для обмена открытыми опубликованными и широко распространяемыми сообщениями («новостями») между сетевыми компьютерами. Это также вся совокупность новостей, которыми обмениваются подобным образом. Usenet – важный элемент Интернета, т. к. этот недорогой и легко устанавливаемый протокол обеспечивает работу почты и не требует выделенного канала. Как следствие, именно Usenet является первым приложением Интернета при подключении нового пользователя.

UTF-8, UTF-16, UTF-32 – кодировки Юникода, использующие 8-, 16- и 32-битное кодирование текста соответственно. UTF-8 и UTF-16 – кодировки переменной длины, означающие, что некоторые символы требуют больше, чем 8 или 16 битов. UTF-32 – кодировка фиксированной длины, означающая, что все символы требуют для кодировки 32 бита.

W3C (World-Wide Web Consortium) – Консорциум Всемирной сети, осуществляющий контроль за развитием протоколов, языков разметки и других технических стандартов Всемирной сети.

XML (Extensible Markup Language) – расширяемый язык разметки – определение языка разметки, который является упрощенной версией SGML и создан как более гибкий, чем HTML, формат доставки информации по Сети, т. к. позволяет определить различные типы разметки. На сегодняшний день языки разметки, определенные в XML, включают языки для веб-контента (XHTML), графики (Scalable Vector Graphics, SVG), математических уравнений (MathML), музыки (MML, MusicML) и многих других приложений.

Список литературы

- Adam, A. 1998. *Artificial Knowing: Gender & the Thinking Machine*. London: Routledge.
- Androutsopoulos, J. 1998. Orthographic variation in Greek e-mails: a first approach. *Glossa* 46, pp. 49–67.
- Anis, J. 1997. A Linguistic Approach to Programming. Arob@se, 1.2. <http://www.liane.net/arobase>
- Barrera-Bassols, N. and Zinck, J.A. 2002. Ethnopedological research: a worldwide review. In 17th World congress of soil science CD-ROM proceedings: Confronting new realities in the 21st century. 590.1-590.12. Bangkok: Kasetsart University. (http://www.itc.nl/library/Papers/arti_conf_pr/barrera.pdf)
- Block, D. 2004. Globalization, transnational communication and the Internet. *International Journal on Multicultural Societies*, Vol. 6, No.1, pp.13–28.
- Climent, S., J. Moré, A. Oliver, M Salvatierra, I Sanchez, M. Taulé and L. Vallmanya. 2004. Bilingual Newsgroups in Catalonia: A Challenge for Machine Translation. *Journal of Computer-Mediated Communication* Vol. 9, No. 1. <http://www.ascusc.org/jcmc/>
- Crystal, D. 2000. *Language Death*. Cambridge: Cambridge University Press.
- . 2001. *Language and the Internet*. Cambridge: Cambridge University Press.
- . 2003. *English as a Global Language, Second Edition*. Cambridge: Cambridge University Press.
- Dalby, A. 2003. *Language in Danger*. New York: Columbia University Press.
- Durham, M. 2004. Language Choice on a Swiss Mailing List. *Journal of Computer-Mediated Communication* 9.1. <http://www.ascusc.org/jcmc/>
- Dunker, E. 2002. Cross-cultural usability of the library metaphor. Proceedings of the second ACM/IEEE-CS joint conference on Digital libraries. Portland, OR.
- Fellbaum, C., and G. Miller. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.
- Ferguson, C. A. 1959. Diglossia. *Word*, 15, pp. 325–340.
- Friedman, B. and H. Nissenbaum. 1995. Minimizing bias in computer systems. Conference companion on Human factors in computing systems, 444. ACM Press.
- Friedman, B. and H. Nissenbaum. 1997. Bias in computer systems. In Friedman, B., ed. *Human Values and the Design of Computer Technology*, pp. 21–40. Stanford, California. Cambridge ; New York, CSLI Publications; Cambridge University Press.

- , 1997. Self-presentation and interactional alliances in e-mail discourse: the style- and codeswitches of Greek messages, *International Journal of Applied Linguistics* 7: pp. 141–164.
- Georgakopolou, A. (Forthcoming). On for drinkies? E-mail cues of participant alignments. In S. Herring (ed.), *Computer-Mediated Conversation*.
- Global Reach. 1999-2005. Global internet statistics by language. Online marketing information. <http://global-reach.biz/globstats/index.php3>
- Greenberg, J. 1956. The measurement of linguistic diversity. *Language*, Vol. 32, No. 2, pp. 109–115.
- Grefenstette, Gregory. 1999. The WWW as a resource for example-based MT tasks. Paper presented at ASLIB “Translating and the Computer” conference, London.
- Grimes, J. E. 1986. Area norms of language size. In B.F. Elson, ed., *Language in global perspective: Papers in honor of the 50th anniversary of the Summer Institute of Linguistics, 1935–1985*, pp. 5–19. Dallas: Summer Institute of Linguistics.
- Hafner, K., and Lyon, M. 1996. *Where Wizards Stay Up Late: The Origins of the Internet*. New York: Simon and Schuster.
- Hård af Segerstad, Y. 2002. Effects of Mobile Text Messaging on Swedish Written Language — human adaptability made visible. *International Conference on Cultural Attitudes towards Technology and Communication, The Net(s) of Power: Language, Culture and Technology*, Montréal.
- Holmes, H. K. 2004. An analysis of the language repertoires of students in higher education and their language choices on the Internet (Ukraine, Poland, Macedonia, Italy, France, Tanzania, Oman and Indonesia). *International Journal on Multicultural Societies*, Vol. 6, No. 1, pp. 29–52.
- Ifrah, G. 1999. *The Universal History of Numbers: From Prehistory to the Invention of the Computer*. New York: John Wiley and Sons.
- Information Sciences Institute. 2003. USC Researchers Build Machine Translation System — and More — For Hindi in Less Than a Month. <http://www.usc.edu/isinews/stories/98.html>
- Kay, Martin, Jean-Mark Gawron, and Peter Norvig. 1993. *Verbmobil: A Translation System for Face-to-Face Dialog*. Stanford, CA: CSLI Publications.
- Krauss, Michael. 1992. The world’s languages in crisis. *Language* Vol. 68, No. 1, pp. 4–10.
- Koutsogiannis, D., and B. Mitsikopolou. 2004. Greeklish and Greekness: Trends and Discourses of “Glocalness”. *Journal of Computer-Mediated Communication* 9.1. <http://www.ascusc.org/jcmc/>

- Lavoie, B. F. and E. T. O'Neill. 1999. How "World Wide" is the Web? *Annual Review of OCLC Research* 1999. 2003.
- Lévénéz, Eric. 2003. Computer languages timeline. <http://www.levenez.com/lang/>
- Lieberson, S. 1964. An extension of Greenberg's linguistic diversity measures. *Language*, 40, pp. 526-531.
- Mafu, S. 2004. From oral tradition to the information era: The case of Tanzania. *International Journal on Multicultural Societies*, Vol.6, No.1, pp. 53-78.
- Muhlhäusler, P. 1996. *Linguistic Ecology: Language Change & Linguistic Imperialism in the Pacific Rim*. London: Routledge.
- Nettle, D. 1999. *Linguistic Diversity*. Oxford: Oxford University Press.
- Nettle, D., and S. Romaine. 2000. *Vanishing Voices: The Extinction of the World's Languages*. Oxford: Oxford University Press.
- Nunberg, Geoffrey. 1998. Languages in the Wired World. Paper presented at La politique de la langue et la formation des nations modernes, Centre d'Etudes et Recherches Internationales de Paris.
- O'Neill, Edward T, Brian F. Lavoie, and Rick Bennett. 2003. Trends in the Evolution of the Public Web: 1998 - 2002. *D-Lib Magazine*, 9.4. <http://www.dlib.org/dlib/april03/lavoie/04lavoie.html>
- O'Neil, E.T.; P.D. McClain; and B.F. Lavoie 1997. A methodology for sampling the World-Wide Web. Technical report, OCLC Annual Review of Research. <http://www.oclc.org/oclc/research/publications/review97/oneill/o'neilla%r980213.html>
- Paolillo, J. C. 1996. Language Choice on soc.culture.Punjab. *Electronic Journal of Communication/Revue Electronique de Communication*, 6(3). <http://www.cios.org/>
- Paolillo, J. C. 2001. Language Variation in the Virtual Speech Community: A Social Network Approach. *Journal of Sociolinguistics*, 5.2.
- Paolillo, J. C. 2002. Finite-state transliteration of South Asian text encodings. In *Recent Advances in Natural Language Processing: Proceedings of the ICON International Conference on Natural Language Processing*. New Delhi: Vikas Publishing House, Ltd.
- Paolillo, J. C. To appear, 2006. 'Conversational' code switching on Usenet and Internet Relay Chat. To appear in S. Herring, ed., *Computer-Mediated Conversation*. Cresskill, NJ: Hampton Press.

- Peel, R. 2004. The Internet and language use: A case study in the United Arab Emirates. *International Journal on Multicultural Societies*, Vol. 6, No. 1, pp. 79–91.
- Phillipson, R. 1992. *Linguistic Imperialism*. Oxford: Oxford University Press.
- Phillipson, R. 2003. *English-Only Europe?* London: Routledge.
- Pimienta, D.; and B. Lamey. 2001. Lengua espacola y cultural hispanicas en la Internet: Comparació con el ingles y el frances. II Congreso Internacional de la Lengua Espanola, Valladolid, 16–19 October 2001.
- Pimienta, D.; et al. 2001. L5: The fifth study of languages on the Internet. <http://funredes.org/LC/english/L5/L5tendencies.html>
- Reed, S. L., and D. B. Lenat. 2002. Mapping Ontologies onto Cyc. American Association for Artificial Intelligence. <http://www.aaai.org/>
- Resnik, P. 1999. Mining the Web for Bilingual Text. 37th Annual Meeting of the Association for Computational Linguistics (ACL'99), College Park, Maryland.
- Rheingold, H. 2000. *The Virtual Community: Homesteading on the Electronic Frontier*, revised edition. Cambridge, MA: MIT Press.
- Skutnabb-Kangas, T., and R. Phillipson. 1995. *Linguistic Human Rights: Overcoming Linguistic Discrimination*. Berlin: Mouton de Gruyter.
- Smith, E. A. 2001. On the co-evolution of linguistic, cultural and biological diversity. In L. Maffi, ed. *On Biocultural Diversity*, 95–117. Washington DC: Smithsonian Institution Press.
- Smith, M. 1999. Invisible Crowds in Cyberspace: Measuring and Mapping the Social Structure of USENET. In M. Smith and P. Kollock, eds., *Communities in Cyberspace*. London: Routledge Press.
- Spencer, H. and Lawrence, D. 1998. *Managing Usenet*. Sebastopol, CA: O'Reilly.
- Su, H.-Y. 2004. The Multilingual and Multi-Orthographic Taiwan-Based Internet: Creative Uses of Writing Systems on College-Affiliated BBSs. *Journal of Computer-mediated Communication* 9.1. <http://www.ascusc.org/jcmc/>
- Torres i Vilatarsana, Marta. 2001. Funciones pragmáticas de los emoticonos en los chats. *Interlingüística* 11.
- Torres i Vilatarsana, Marta. 1999. Els xats: entre l'oralitat i l'escriptura. Article publicat a la revista *Els Marges*, 65 (desembre, 1999). Publicat a Internet (gener, 2001) amb el consentiment d'aquesta revista.

- UNESCO. 2003. Cultural and Linguistic Diversity in the Information Society. UNESCO publications for the World Summit on the Information Society. CI.2003/WS/07 <http://unesdoc>
- [UNESCO.org/images/0013/001329/132965e.pdf](http://unesco.org/images/0013/001329/132965e.pdf)
- Unicode Consortium. 1991. The Unicode Standard: Worldwide Character Encoding. Reading, Mass., Addison-Wesley Pub.
- Unicode Consortium. 1996. The Unicode Standard, Version 2.0. Reading, Mass., Addison-Wesley Developers Press.
- Unicode Consortium. 2000. The Unicode Standard, Version 3.0. Reading, Mass., Addison-Wesley.
- Unicode Consortium. 2003. The Unicode Standard, Version 4.0. Reading, Mass., Addison-Wesley.
- Warschauer, M., G. R. El Said and A. Zohry. 2002. Language Choice Online: Globalization and Identity in Egypt. *Journal of Computer-Mediated Communication (JCMC)*, 7.4. <http://www.ascusc.org/jcmc/>
- Wasserman, Herman. 2002. Between the local and the global: South African languages and the Internet. Litnet Seminar Room. <http://www.litnet.co.za/seminarroom/11wasserman.asp>
- Wright, S. 2004. Introduction. *International Journal on Multicultural Societies*, Vol. 6, No. 1, pp. 3–11.
- Wurm, S. A.. 1991. Language death and disappearance: causes and circumstances. In R. H. Robbins and E. M. Uhlenbeck, eds., *Endangered Languages*, 1–18. Oxford: Berg.
- Wurm, S. A., ed. 1996. *Atlas of the World's Languages in Danger of Disappearing*. Paris: UNESCO Publishing/Pacific Linguistics.