Альтернативные точки зрения

Языковое разнообразие в Интернете: ситуация в Азии

Йошики Миками*, Ахамед Заки абу Бакар●, Вираш Сонлерт-ламваниш○, Ом Викас■, Заварски Павол*, Мохд Заиди Абдул Розан*, Гендри Надь Янош▲, Томоэ Такахаши*

(Члены Проекта «Обсерватория языков» (Language Observatory Project, LOP), Японское агентство по науке и технологии)

«Прежде чем закончить это письмо, я хочу довести до сведения Вашего Преосвященства тот факт, что в течение многих лет я жаждал увидеть в данной Провинции какие-нибудь книги, напечатанные на языке этой страны и на ее алфавите, какие видел я в Малабаре к большой чести тамошней христианской общины. Сделать это мне не удалось по двум причинам: первая из них в том, что казалось

* Технологический университет г. Нагаока (Nagaoka University of Technology), Япония; ● Малазийский технологический университет (Universiti Tekmologi Malaysia), Малайзия; ○ Тайская лаборатория вычислительной лингвистики (Thai Computational Linguistic Laboratory), Таиланд; ■ Технологический департамент индийских языков, Министерство информационных технологий (Technology Department of Indian Languages, Ministry of Information Technology), Индия; ▲ Университет г. Мишкольц (Miskolc University), Венгрия.

Адрес для контактов с авторами: mi-kami@kjs.nagaokaut.ac.jp

невозможным составить текст из такого количества форм, число которых доходило до шести сотен против наших двадцати четырех в Европе...» Письмо отца-иезуита Фриара в Рим, 1608 (Priolkar, 1958).

«Когда 500 лет назад в Майнце Гуттенберг напечатал свою знаменитую Библию, ему нужно было только одно основное клише для каждой буквы алфавита, а в 1849 году, когда издательство American Mission Press напечатало Библию в Бейруте на арабском, то использовало не менее 900 знаков, и даже такого большого числа оказалось недостаточно...» Джон М. Мунро (John M. Munro), 1981 (Lunde, 1981).

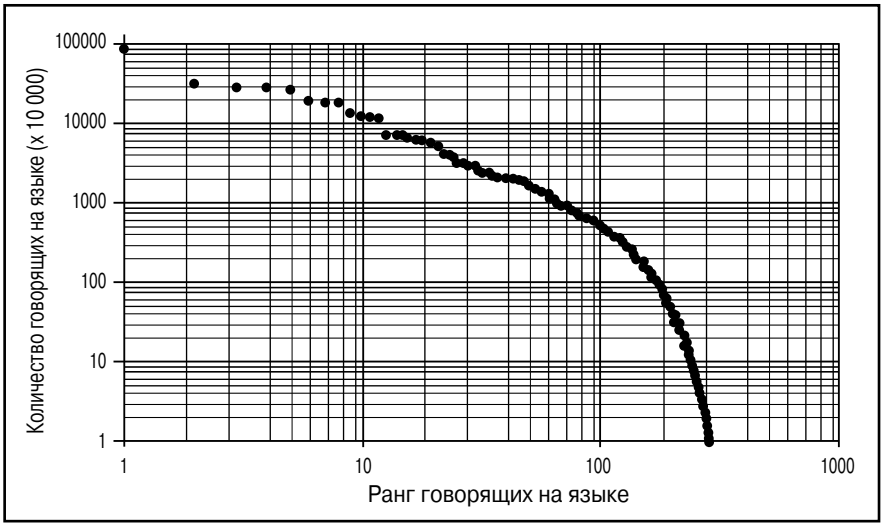
Разнообразие языков и алфавитов в Азии

По оценкам специалистов на сегодняшний день в мире существует 7000 устных языков (Gordon, 2005). Что касается официальных языков, то их по-прежнему много и, может быть, их количество превышает три сотни. Текст универсального значения – Всеобщая декларация прав человека – был переведен Управлением Верховного комиссара ООН по правам человека (United Nations Higher Commission for Human Rights, UNHCHR) на 328 языков (UNHCHR, 2005).

Из всех языков, представленных на сайте этой организации, самую большую аудиторию имеет китайский язык – почти миллиард человек, за ним идут английский, русский, арабский, испанский, бенгали, хинди, португальский, индонезийский и японский. В конце этого списка стоят языки, на которых говорит менее ста тысяч человек. Азиатские языки занимают шесть из 10 верхних позиций и почти половину (48) из первых по распространенности 100 языков.

На сайте Управлением Верховного комиссара ООН по правам человека (УВКПЧ) представлены также оценочные данные по количеству людей, говорящих на каждом языке. Если мы ранжируем языки по количеству говорящих на них людей и нанесем языки на график, построенный в логарифмическом масштабе, то увидим, что соотношение количества говорящих на языке и их ранга (позиции) среди говорящих на других языках мира практически соответствует кривой закона Ципфа (Рис. 1) (по крайней мере, в интервале от десятков до сотен).

Рис. 1 Кривая квази-закона Ципфа



Языковое разнообразие в Азии становится более явным, если посмотреть на разнообразие алфавитов, используемых для представления языков. С позиции сложности локализации разнообразие алфавитов – это проблема. Трудно ответить на вопрос, сколько алфавитов существует в мире, поскольку ответ зависит от единицы измерения. В данной статье для простоты мы объединяем в одну категорию все латинские алфавиты, алфавиты и их расширения для европейских языков, вьетнамского, филиппинского и др. Мы принимаем за одну категорию кириллические и арабские языки. Точно так же, в рамках одной категории мы рассматриваем китайские иероглифы, японское силлабическое письмо и корейский хангыль. Остальные алфавиты включают индийские письменности, которые составляют пятую категорию. В нее входят не только индийские алфавиты типа деванагари, бенгали, тамильский, гуджарати и другие, но и 4 крупнейших языка Юго-Восточной Азии: тайский, лаосский, камбоджийский (кхмерский) и мьянмский. Невзирая на разницу в написании, эти алфавиты имеют общее происхождение (древний язык брахми) и ведут себя одинаково при кодировании. Если сложить число людей, говорящих на каждом языке, в соответствии с данной группировкой по алфавитам, то мы получим картину, представленную в Таблице 1. Тогда алфавиты, используемые в Азии, распространятся на все пять категорий, в то время как алфавиты, ис-

пользуемые в других частях мира, представляют собой, в основном, латинский, кириллический, арабский и некоторые другие.

Таблица 1. Распределение групп пользователей по основным категориям алфавитов

Алфавит	Латинский	Кириллический	Арабский	Ханьцзы	Индийская группа	Другие*
Кол-во пользователей (млн)	2 238	451	462	1 085	807	129
% от общего числа	43,28	8,71	8,93	20,98	15,61	2,49

* Другие: греческий, грузинский, армянский, амхарский, дивехи, иврит и пр.

Современное состояние охвата языков – ситуация с Windows

По сравнению с прошедшим десятилетием современные продукты информационно-телекоммуникационных технологий (ИКТ) способны до известной степени оперировать с многоязычием. Благодаря появлению стандарта кода многоязычных знаков в виде ISO/IEC 10646, который используется и для стандарта Юникод, а также благодаря сложной интернационализации программного обеспечения в течение 10 лет выросло количество языков, поддерживаемых основными настольными платформами ИКТ. Однако охват языков данными платформами все еще ограничен. Самая последняя версия Windows XP (Professional SP2)* способна работать с 123 языками. Однако, если мы внимательнее посмотрим на этот список, то увидим, что большинство представленных в нем языков – европейские и лишь немного азиатских и африканских языков. Охват языков показан в Таблице 2. Здесь языки разбиты по группам алфавитов, как это было описано в первой части статьи. Так, по подсчетам, на тех языках, с которыми работает Windows XP, говорят 83,72 % от общего числа населения Земли. Хотя данная цифра

* По состоянию на год издания оригинальной версии данного сборника.

может быть истолкована как довольно неплохая, нам она кажется завышенной, и плохо соотносится с реальностью, что мы покажем ниже.

Таблица 2. Охват языков Windows XP SP 2 с разбивкой по основным категориям алфавитов

Регион алфавита	Латинский	Кириллица	Арабский	Ханьцзы	Индийская группа	Прочие
Европа	Европейские* и славянские** языки	Русский, македонский и славянские языки***	–	–	–	Греция Грузия Армения
Азия	Азербайджанский, вьетнамский, малайский, индонезийский, узбекский, турецкий	Монгольский, азербайджанский, казахский, киргизский, узбекский	Арабский, урду, персидский	Китайский, японский, корейский	Гуджарати, тамильский, телугу, каннада, бенгальский, малаялам, пенджаби, хинди, маратхи, санскрит, конкани, ория, тайский	Ассирийский, дживехи, иврит

* Включают: албанский, баскский, каталанский, датский, голландский, английский, эстонский, фарерский, финский, французский, галисийский, немецкий, венгерский, исландский, итальянский, латвийский, литовский, мальтийский, норвежский, португальский, румынский, саами, испанский, шведский и валлийский языки.

** Включают: сербский, чешский, хорватский, словацкий, боснийский, польский и словенский языки.

*** Включают: белорусский, болгарский, сербский, боснийский и украинский языки.

Ситуация с Google

Поисковые машины стали неотъемлемой составляющей глобального информационного общества. Их работа делает доступным огромный массив знаний. Когда мы изучаем охват языков наиболее распространенными поисковыми машинами, мы видим, что ситуация здесь гораздо хуже, чем в случае с Windows. Одна из самых распространенных многоязычных поисковых машин – Google – проиндексировала по состоянию на апрель 2005 г. свыше 8 млрд страниц на разных языках мира. Однако оказалось, что эти страницы представляют всего лишь 35 языков. Среди них только 7 азиатских языков: индонезийский, арабский, китайский традиционный, китайский упрощенный, японский, корейский и иврит (Таблица 3). Если подсчитать численность охваченного населения, то она сократится до 61,37 % именно потому, что материалы, написанные на азиатских и африканских языках, недоступны для поиска.

Таблица 3. Охват языков Google с разбивкой по основным категориям алфавитов

Регион алфавита	Латинский	Кириллица	Арабский	Ханьцзы	Индийская группа	Прочие
Европа	Европейские* и славянские** языки	Русский, болгарский, сербский	–	–	–	Греция
Азия	Индонезийский	–	Арабский	Китайский традиционный и упрощенный, японский, корейский	–	Иврит, турецкий

*Включает: каталанский, датский, голландский, английский, эстонский, финский, французский, немецкий, венгерский, исландский, итальянский, латвийский, литовский, норвежский, португальский, румынский, испанский и шведский языки.

**Включает: хорватский, чешский, польский, словацкий и словенский языки.

Многоязычный характер Всеобщей декларации прав человека

Приведем еще один пример. Как мы упомянули в начале нашей статьи, на веб-сайте Управления Верховного комиссара ООН по правам человека Всеобщая декларация прав человека представлена на более чем 300 языках мира, начиная с абхазского и заканчивая зулу. К сожалению, там также можно найти и много переводов на разные языки, особенно, на языки с нелатинским алфавитом, и эти переводы даны в виде GIF- или PDF-файлов, а не в виде кодированных текстов. И снова, как и в предыдущих случаях, представим эту ситуацию в виде таблицы (Таблица 4). Из нее понятно, что в виде кодированных текстов лучше всего представлены языки, использующие латинский алфавит. Языки, использующие другие алфавиты, индийские в особенности, с трудом поддаются кодировке. Если алфавит не удается представить в какой-то одной из имеющих трех форм, он попадает в категорию «недоступных». Более того, не так просто загрузить специальные шрифты для надлежащего просмотра этих алфавитов. Сложность ситуации можно назвать цифровым разрывом между языками или «языковым цифровым разрывом».

Таблица 4. Представление Всеобщей декларации прав человека с разбивкой по основным категориям алфавитов

Регион алфавита	Латинский	Кириллица	Арабский	Ханьцзы	Индийская группа	Прочие
Европа	Европейские и славянские языки	Русский, болгарский, сербский	–	–	–	Греция
Азия	Индонезийский	–	Арабский	Китайский традиционный и упрощенный, Японский, корейский	–	Иврит, турецкий

Регион алфавита	Латинский	Кирил-лица	Арабский	Ханьцзы	Индий-ская группа	Прочие
В каком алфавите представ-лен	Латинский	Кирил-лица	Арабский	Ханьцзы	Индийский	Другие
Кодировка	253	10	1	3	0	1
PDF	2	4	2	0	7	10
GIF	1	3	7	0	12	7
Недо-ступны	0	0	0	0	1*	1*

* Недоступными языками являются магади и бходжпури.

Локализация информационных технологий – взгляд в прошлое

Давайте посмотрим, что было 500 лет назад, когда была изобретена эпохальная технология книгопечатания. Буквопечатающая технология была независимо изобретена и на Востоке, и на Западе. На Востоке эта технология была впервые создана в XIII веке корейскими ремесленниками и затем подхвачена китайцами. Но технология эта не получила развития и впоследствии была вытеснена ксилографией. Буквопечатающая технология, распространенная сегодня повсеместно в Азии, уходит своими корнями в изобретение, сделанное Гуттенбергом в середине XV века.

Первый печатный пресс был привезен на Гоа в 1556 г. Считается, что это – первая печатная машина, привезенная в Азию. Вслед за ней другие машины были привезены в Манилу, Малакку, Макао и другие города Азии. Поначалу эти машины использовались для печати переводных или транслитерированных священных текстов с применением латинского алфавита, но позднее на них стали печатать разные тексты с отпечатками букв местных алфавитов. По мнению одного индийского историка первым печатным текстом в Азии с использованием местного алфавита стала книга на тамильском языке «Христианская доктрина».

На второй странице этого текста содержится рассказ о том, какой подход был использован при локализации печатной технологии на тамильском языке. Несмотря на то, что в тамильском языке всего 246 слогов, на второй странице представлено более 150 знаков в комплекте шрифта. Отец-иезуит, проживавший в XVII веке где-то на берегу Малабара, писал в Рим: «…в течение многих лет я жаждал увидеть в этой Провинции какие-нибудь книги, напечатанные на языке этой страны и на ее алфавите, …но сделать это мне пока не удалось. Главная причина в том, что мы должны составлять текст из более чем 600 отпечатков против 24 в Риме…» (Priolkar, 1958).

В Маниле, в то время центре испанских колоний, «Доктрина» была переведена на тагалогский язык в 1593 г. Однако так случилось, что перевод сопроводили транслитерированным текстом. Тагалогская версия «Доктрины» была составлена в трех вариантах: на тагалогском языке с использованием тагалогского алфавита; на тагалогском языке с использованием латинского алфавита и на испанском языке с использованием латинского алфавита. За последующие 100 лет после того, как буквопечатающая технология была привезена в Манилу, два вторые варианта полностью вытеснили первый. В итоге тагалогский алфавит был полностью забыт даже местным населением (Hernandez, 1996). Изображение тагалогского шрифта на почтовой марке, выпущенной почтой Филиппин в 1995 г., показывает нам, как выглядел этот шрифт, и служит напоминанием об утерянном культурном наследии.

Эти два исторических события дают нам урок: когда локализация реализуется неудачно, появление новой технологии может разрушить систему письменности или даже саму культуру.

Стандарты кодировки как краеугольный камень локализации

За цифровым языковым разрывом стоит множество причин: экономических, политических, социальных и пр. Однако с технических позиций локализация должна оставаться главным фактором. Как явствует из письма отца-иезуита, отправленного 400 лет назад в Рим, отрывок из которого мы привели в самом начале нашей статьи, даже во времена технологии книгопечатания пионеры информационных технологий были вынуждены преодолевать аналогичные сложности, локализуя технологии в другую языковую среду, почти так же, как это делают сего-

дня инженеры-компьютерщики. Особым препятствием для нелатинских алфавитов является отсутствие или недоступность соответствующих стандартов кодировки. По этой причине разработчикам веб-сайта с текстом Всемирной декларации прав человека пришлось поместить файлы, не поддающиеся кодированию, в виде изображений или в формате PDF. Если мы посмотрим на международно-признанные справочники схем кодирования, например, IANA Registry of Character Codes (IANA, 2005) или ISO International Registry of Escape Sequences (IPSJ/ITSCJ, 2004), то не сможем найти в них схемы кодирования для таких языков, которые считаются «упущенными сквозь ячейки сети». Следует отметить, что многие стандарты кодировки, принятые на национальном уровне, используются для нескольких языков и имеют национальный статус. Для семьи индийских письменностей первый национальный индийский стандарт был принят в 1983 г. Он получил название Indian Standard Script Code for the Information Interchange (ISSCII). Позже, в 1991 г., он претерпел изменения и вышел во втором издании под названием «National Standard IS 13194», который и используется в Индии в настоящий момент. Однако, несмотря на существование национальных стандартов, поставщики технических средств, разработчики шрифтов и даже конечные пользователи сами создавали собственные таблицы кодирования, что приводило к неразберихе. Стимулом для создания так называемых экзотических схем кодирования или локальных внутренних кодировок послужило внедрение дружественных для пользователя средств создания шрифтов. Несмотря на то, что прикладные системы для этой области не являются автономными и широко распространяются через Сеть, необходимость в стандартизации не была осознана пользователями, поставщиками или разработчиками шрифтов. Отсутствие профессиональных ассоциаций и соответствующих государственных учреждений – еще одна причина сложившейся неконтролируемой ситуации. Интересное исследование по всему многообразию индийских языков провела компания Aruna Rohra and Ananda of Saora Inc. (www.gse.uci.edu/markw/languages.html): на 49 тамильских веб-сайтах она обнаружила существование 15 различных схем кодирования (Aruna & Ananda, 2005).

UCS/Unicode

Первая версия Универсального многооктетного набора кодированных символов (Universal Multiple-Octet Coded Character Set, UCS, ISO/IEC

10646) была выпущена в 1993 г. Юникод, разработанный изначально промышленным консорциумом, приведен сегодня в соответствие последней версии UCS и мог бы устранить неразбериху. Но он не стал доминирующим, по крайней мере, в азиатской части мира. Наше последнее исследование показало, что код UTF-8 охватывает только 8,35 % всех веб-страниц под азиатским ccTLD (Mikami, et al., 2005). Первые и последние десять ccTLD показаны в Таблице 5. Несмотря на то, что ожидается высокая скорость миграции, процесс этот следует тщательно отслеживать.

Таблица 5. Доля веб-страниц, использующих UTF-8, по ccTLD

ccTLD	Название	Доля	ccTLD	Название	Доля
tj	Tajikistan Таджикистан	92,75 %	uz	Uzbekistan Узбекистан	0,00 %
vn	Viet Nam Вьетнам	72,58 %	tm	Turkmenistan Туркменистан	0,00 %
np	Nepal Непал	70,33 %	sy	Syria Сирия	0,00 %
ir	Iran Иран	51,30 %	mv	Maldives Мальдивы	0,00 %
tp	Timor East Восточный Тимор	49,40 %	la	Lao Лаос	0,01 %
bd	Bangladesh Бангладеш	46,54 %	ye	Yemen Йемен	0,05 %
kw	Kuwait Кувейт	36,82 %	mm	Myanmar Мьянма	0,07 %
ae	UAE ОАЭ	35,66 %	ps	Palestine Палестина	0,12 %
lk	Sri Lanka Шри Ланка	34,79 %	bn	Brunei Бруней	0,36 %
ph	Philippines Филиппины	20,72 %	kg	Kyrgyzstan Киргизстан	0,37 %

Источник: Language Observatory Project.

Цели проекта «Обсерватория языков»

Проект «Обсерватория языков» был запущен в 2003 г. как признание растущего значения мониторинга уровня языковой активности в киберпространстве (Language Observatory Project, LOP; UNESCO, 2004). Проект призван создать средства для оценки уровня использования каждого языка в киберпространстве. Если говорить точнее, то от проекта ждут периодического предоставления статистики по языкам, алфавитам и кодировкам в киберпространстве. После полного запуска проект должен будет дать ответы на следующие вопросы:

- Сколько разных языков существует в виртуальной вселенной?
- Какие языки отсутствуют в виртуальной вселенной?
- Сколько веб-страниц написано на определенном языке, скажем, на пушту?
- Сколько веб-страниц написано с использованием тамильского варианта письма?
- Какие виды схем кодирования символов используются для кодирования какого-то определенного языка, скажем, берберского?
- Как быстро Юникод замещает традиционные и локальные схемы кодирования в сети?

Наряду с таким анализом проект будет заниматься подготовкой предложений по преодолению сложившейся ситуации, как на техническом, так и на политическом уровнях.

Альянс проектов

В настоящее время несколько групп экспертов работают в тесном сотрудничестве в рамках обсерватории языков. Организациями-учредителями проекта являются: Технологический университет г. Нагаока (Nagaoka University of Technology), Япония; Токийский университет зарубежных исследований (Tokyo University of Foreign Studies), Япония; Университет Кейо (Keio University), Япония; Технологический университет Малайзии (Universiti Teknologi Malaysia), Малайзия; Университет г. Мишкольц (Miskolc University), Венгрия; проект «Технологическое

развитие индийских языков» (Technology Development of Indian Languages) под руководством Министерства информационных технологий Индии; Лаборатория исследований в области коммуникации (Communication Research Laboratory), Таиланд. Финансирование проекта осуществляется Агентством по науке и технологиям Японии (Japan Science and Technology Agency) в рамках программы RISTEX (RISTEX, 2005). ЮНЕСКО выразила официальную поддержку данному проекту с самого начала его создания. Основные технические компоненты «Обсерватории языков» включают мощную технологию поиска в Сети (кролер-технология) и технологию идентификации особенностей языков (Suzuki, et al., 2002). В проекте используется UbiCrawler (Boldi, et al., 2004) – масштабируемый, полностью распределенный веб-кролер, разработанный совместно Отделом информационных наук Исследовательского унивеситета Милана (Dipartimento di Scienze dell'Informazione, Universita degli Studi di Milano) и Институтом информатики и телематики Итальянского национального совета по исследованиям (Instituto di Informatica e Telematica). Это мощная машина по сбору данных для «Обсерватории языков». Краткое описание совместных усилий проекта и команды UbiCrawler можно найти в публикации ЮНЕСКО (UNESCO, 2004).

Заключение

В данной статье мы стремились подчеркнуть значение мониторинга поведения и активности мировых языков в киберпространстве. Проект «Обсерватория языков» предусматривает использование сложных научных методов для понимания и мониторинга мировых языков. Консорциум проекта надеется, что ему удастся сделать так, чтобы мир больше знал о живущих и умирающих языках. В этом случае можно будет предпринять шаги, чтобы предотвратить исчезновение языков, оказавшихся в тяжелой ситуации. Чтобы эти усилия принесли плоды, «Обсерватория» должна стать центром развития человеческого капитала и депозитарием языковых ресурсов. Накопление цифровых языковых ресурсов в результате проведенных научно-исследовательских работ позволит развивающимся странам и региональным сообществам вывести свои языки в киберпространство и, тем самым, спасти национальное наследие от исчезновения.

Список литературы

Aruna, R. & Ananda, P. 2005. Collecting Language Corpora: Indian Languages. The Second Language Observatory Work Shop Proceedings. Tokyo University of Foreign Studies, Tokyo.

Boldi, P., Codenotti, B., Santini, M., & Vigna, S. 2004. UbiCrawler: A scalable fully distributed Web crawler. Software: Practice & Experience, Vol. 34, No. 8, pp. 711–726.

Gordon, R. 2005. Ethnologue: Languages of the World 15th Edition. (http://www.ethnologue.com/).

Hernandez, Vincente S. 1996. History of Books and Libraries in the Philippines: Manila, The National Commission for Culture and the Arts, pp. 24–31.

IANA. 2005. Character Sets. (http://www.iana.org/assignments/character-sets).

IPSJ/ITSCJ. 2004. International Register of Coded Character Sets to be used with Escape Sequences. (http://www.itscj.ipsj.or.jp/ISO-IR/).

Mikami, Y., Zavarsky, P., Zaidi, M., Rozan, A., Suzuki, I., Takahashi, M., Maki, T., Ayob, I. N., Boldi, P., Santini, M. & Vigna, S. 2005. The Language Observatory Project (LOP). Proceedings of the Fourteenth International World Wide Web Conference, May 2005. Chiba, Japan,. pp. 990Ц991.

Lunde. P. 1981. Arabic and the Art of Printing. Saudi, Aramco World.

Priolkar, A. K. 1958. The Printing Press in India – Its Beginning and Early Development. Bombay, Marathi Samshodhana Mandala, pp. 13–14.

RISTEX. 2005. (http://www.ristex.jp/english/top_e.html).

Suzuki, I., Mikami, Y., Ohsato, A. & Chubachi, Y. 2002. A language and character set determination method based on N-gram statistics, ACM Transactions on Asian Language Information Processing, Vol. 1, No. 3, pp. 270–279.

UNESCO. 2004. Parcourir le cyberespace á la recherche de la diversité linguistique. UNESCO WebWorld News, 23rd Feb. 2004. (http://portal.UNESCO.org/ci/en/ev.php-URL_ID=14480&URL_DO=DO_TOPIC&URL_SECTION=201.html).

UNHCHR. 2005. Universal Declaration of Human Rights. (http://www.unhchr.ch/udhr/navigate/alpha.htm).