Supporting Access to Consumer Health Information Across Languages

Graciela Rosemblat, Tony Tse, Darren Gemoets, John E. Gillen, Nicholas C. Ide
Lister Hill National Center for Biomedical Communications,
National Library of Medicine, NIH, DHHS, Bethesda, MD

ABSTRACT

Access to high-quality online consumer health information is frequently a barrier for non-English speakers and speakers of English as a foreign language. One strategy is to leverage existing English-language systems. In this case study, we present mechanisms developed, designed, and implemented to support access by Spanish speakers to ClinicalTrials.gov. First, we briefly introduce the underlying cross-language information retrieval (CLIR) technology and development of a Spanish-English bilingual term list. We then focus on language-based design decisions and their implications for integrating components of a Spanish-language system with the existing ClinicalTrials.gov system architecture. Data display and navigation issues will also be presented. We conclude with results from a preliminary evaluation of the Spanish-English CLIR prototype. The long-term goal is to adapt such technologies generally for consumer health sites to support access in different languages.

Key words: cross-language information retrieval; bilingual term list, consumer health information system

INTRODUCTION

Language is often a barrier in accessing online health information for non-native English speakers. Not only is a considerable amount of consumer health information written in English, but also obtaining it online requires the use of English-language Web sites, navigation tools, and information retrieval systems. Although the number of health-related documents in other languages is increasing, evaluation results for health sites in Spanish in particular indicate a marked disparity with the English sites, both in quality and quantity of information provided [1]. Fundamentally, the availability of quality consumer health information in languages other than English is not keeping pace with the rapid expansion of the Web as a global communication tool.

The problem is compounded by the gap between health expressions used by laypersons and technical terms used by healthcare professionals. Consumers typically have difficulty formulating health-related queries and understanding medical terms or jargon, independent of primary language [2]. Thus, developing a general mechanism to support access for speakers of other languages would leverage the numerous existing English-language consumer health information systems. We present an approach for providing partial access to Spanish speakers, through the expansion and reuse of an existing English-language system, ClinicalTrials.gov.

In the consumer health domain, several approaches have been used to address the language barrier. One strategy is to create resources in languages other than English. For example, AIDSinfo [1] creates original Spanish-language materials about HIV/AIDS through human medical translators [3]. Another strategy is to link to health topics already available in Spanish, curated by expert reviewers and editors, such as the New York Online Access to Health (NOAH)[2] [4] and the U.S. National Library of Medicine (NLM) MedlinePlus en español®[3] Web sites.

Some sites use controlled medical vocabularies available in multiple languages. For example, HONSelect[4] allows consumers to find health information in English and four other languages by using

translations of Medical Subjects Headings (MeSH®) terms [5,6]. After MeSH descriptors are automatically assigned to health information resources, a string representing the medical concept may be displayed in any language. However, documents retrieved at HONSelect are also derived from existing sources that provide information originally written in a particular language.

Although machine translation (MT) technology may yet be a feasible strategy, it is not at present sufficiently precise for the real-time, unedited performance required in health information [7]: "Machine translation is demonstrably cost-effective for large scale and/or rapid translation of (boring) technical documentation, (highly repetitive) software localization manuals. ... The human translator is (and will remain) unrivaled for non-repetitive linguistically sophisticated texts (e.g. in literature and law), and even for one-off texts in specific highly-specialized technical subjects. (p. 2)"

BACKGROUND

In response to the rapid growth of the Spanish-speaking population in the U.S. and their presence online, many health-related organizations now offer online resources in Spanish. For example, NLM launched MedlinePlus en español in 2002. Subsequent NLM focus groups of Hispanic-Americans and online surveys indicated a need for Spanish-language access to its ClinicalTrials.gov[5] Web site, which provides information about clinical research studies for patients and other members of the general public. As a first step in addressing this need, we have adapted cross-language information retrieval (CLIR) techniques to facilitate access by Spanish-speaking users to this English-language Web site (Figure 1) [8,9].

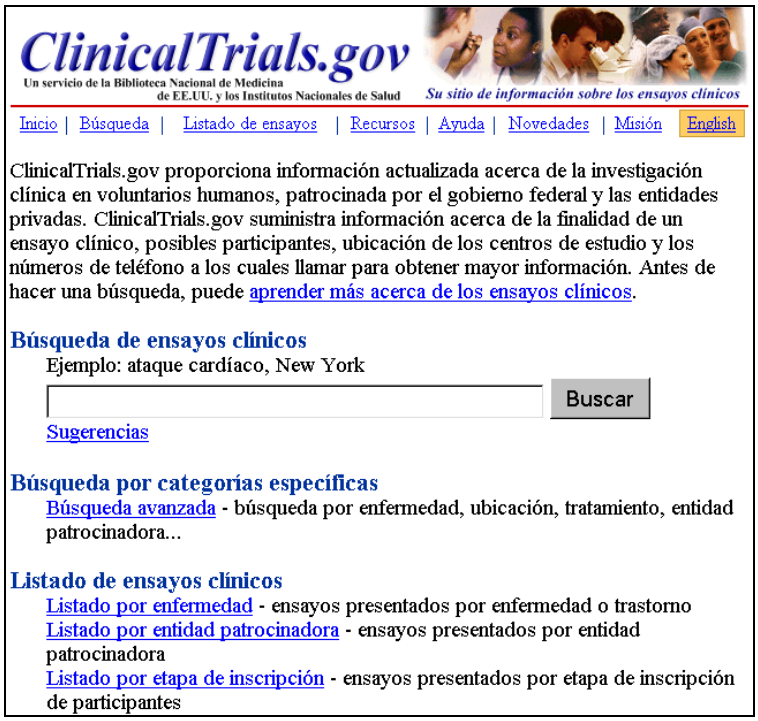


Figure 1. Spanish ClinicalTrials.gov prototype home page

ClinicalTrials.gov presently contains over 14,500 records summarizing studies sponsored by the U.S. National Institutes of Health (NIH), other governmental agencies, and the pharmaceutical industry from around the world. Launched in February 2000, the site's technical infrastructure is constantly being improved, including the usability of its search and display features, based on formative evaluation [10].

Given this well-established, yet dynamic system, it follows that any new components for Spanish-language support would need to be compatible with the current architecture and design of ClinicalTrials.gov, but also have flexibility to evolve. Furthermore, the content and information retrieval (IR) mechanisms in both language sites must be parallel and synchronized to provide timely, accurate, usable results.

CROSS-LANGUAGE INFORMATION RETRIEVAL (CLIR)

Typically, cross-language information retrieval (CLIR) requires translation of either the query or the document so that information can be retrieved. In our earlier study using machine translation (MT) within the ClinicalTrials.gov environment, translation of Spanish queries into English (query translation) outperformed the document translation approach [9]. We believe this to be due to:

- IR features (synonym expansion; lexical variant handling) built into the search engine customized for ClinicalTrials.gov;
- Incoming user queries, typically keyword-based nouns or phrases for treatments and conditions, that do not require part-of-speech or lexical disambiguation; and
- Domain-restricted nature of queries and documents in ClinicalTrials.gov, which minimizes the need for sense disambiguation.

In a subsequent study, we compared two approaches to query translation-based CLIR: a bilingual term list (BTL) versus MT [9]. The BTL, which we created from publicly available sources[6], provided a technologically simpler, more customizable, and system-compatible alternative to a high-quality but proprietary MT system. A BTL is a look-up table where words and expressions in one language (e.g., Spanish) are matched with their equivalents in another language (e.g., English). Other BTL modifications and enhancements include Spanish normalization algorithms for number (singular/plural) and gender (male/female), addition of domain-specific and consumer health terms, and removal of error-filled general vocabularies. Curation of portions of the BTL improved translation quality. As BTL-based CLIR performance improved, our focus turned to comparisons between the Spanish- and English-language systems.

We have continued fine-tuning the quality and size of the BTL by adding new terms or glossaries only after exhaustive line-by-line human review. Maintaining the BTL is an ongoing process, as new terms and novel expressions appear in trial protocols and are translated.

DATA MANAGEMENT

Two types of Web pages are displayed at ClinicalTrials.gov: static and dynamic. Static pages contain relatively stable text, such as information about the site (e.g., “About”) and search tips (e.g., “Basic Search”). In these cases, MT followed by manual expert editing (or post-editing) is feasible. Nonetheless, keeping both language versions of the static pages consistent is a challenge.

In contrast, dynamic pages, such as detailed study data in the records, change frequently (e.g., “Study Summary”), rendering manual post-editing not viable. However, because many fields in each record use controlled vocabulary terms (e.g., “Overall Status”), they only need post-editing once. Such terms could then be displayed with 100% accuracy in Spanish.

The remaining free-text fields contain considerably more content, including technical terms, and change more frequently than controlled fields. Post-editing these fields is resource intensive. Thus, we only translated the “Brief Title,” which is intended to encapsulate the purpose of the study in lay language. By clicking a Spanish explanatory link, users have access to the remaining free-text fields in the original

English, on a separate window. We believe that this limited translation into Spanish in the proposed prototype provides the "gist" of the study, facilitating relevance determination for Spanish speakers (see "Preliminary Evaluation" section). Some considerations included:

- Human resources to carefully post-edit MT results;
- Style and variability in the translation of the same natural language text;
- Handling of proper names: organizational affiliations and location information were not translated in order to avoid confusion, as they frequently lack official Spanish equivalents; and
- Constant review and updates of translations to remain synchronized with the English text.

Based on this last consideration, we modified the system to track new (untranslated) and changed titles in English, and provided a suitable mechanism for entering translations in Spanish with minimal time lag. This mechanism extends the existing Protocol Registry System used for registering new Clinical Trials. Human translators may log into the system, view all new or changed titles, and perform any required edits or translations. These changes are saved and used by ClinicalTrials.gov after the next index.

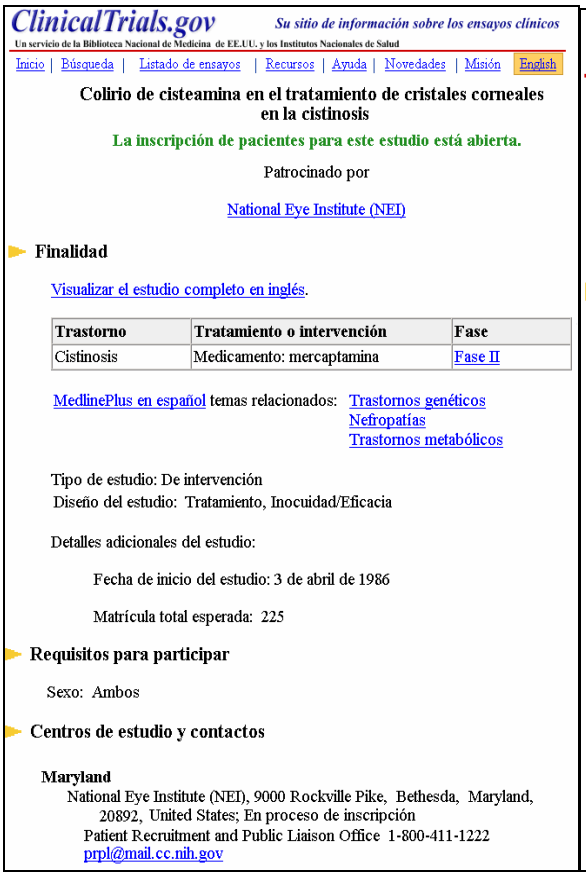
DATA DISPLAY

In general, in a Spanish monolingual system, Spanish-language queries retrieve Spanish-language documents. However, in a Spanish-English CLIR system, queries submitted in Spanish retrieve documents in English. Although some Spanish-speakers are proficient in reading English, we believe that displaying text fully in English is not sufficient for all Spanish-speaking users to derive relevance determination, as many may have varying levels of English proficiency. Thus, we considered other potential solutions to make the information more accessible to all segments of the Spanish-speaking population. Unedited machine translation was ruled out for use in the health domain, where mistranslation may have dangerous outcomes. While the accuracy of human translation (post-editing) is unsurpassed, it is extremely resource intensive and difficult to scale up. Therefore, we devised a system for Spanish display of documents that is accurate, scaleable, and requires few resources.

"Doclets" are abbreviated ClinicalTrials.gov records that display most controlled fields and the manually translated "Brief Title" free-text field (Figures 3.1 and 3.2). The remaining free-text fields (e.g., "Purpose" and "Detailed Description") are accessible in English: clicking links labeled in Spanish allows users to view the corresponding full English-language record in a separate window. Note that some controlled fields, such as center locations and organization names, were *not* translated into Spanish to avoid possible ambiguity (e.g., lack of official translations). This type of information is presented in the original language: French locations and/or centers are presented in French, Italian ones in Italian, and so on. This is also the case in the full English-language records whenever foreign researchers or sponsors provide us with the data in the native language.

Because the focus of ClinicalTrials.gov design has been on usability and support of consumer access to information in the medical domain [10,11], we used the English system as the model for the Spanish-language CLIR-based prototype. Nevertheless, we addressed several language-related issues, in particular:

- Manipulation of the character set for correct diacritics display in all Spanish Web pages;
- Linking to relevant outside Spanish-language resources, when available, such as MedlinePlus en español;
- Modification of the existing Web interface to generate separate screens for Spanish; and
- Selection of language to display at run time.



ClinicalTrials.gov — Su sitio de información sobre los ensayos clínicos

Un servicio de la Biblioteca Nacional de Medicina de EE.UU. y los Institutos Nacionales de Salud

Inicio | Búsqueda | Listado de ensayos | Recursos | Ayuda | Novedades | Misión | English

Colirio de cisteamina en el tratamiento de cristales corneales en la cistinosis

La inscripción de pacientes para este estudio está abierta.

Patrocinado por

National Eye Institute (NEI)

Finalidad

Visualizar el estudio completo en inglés.

Trastorno	Tratamiento o intervención	Fase
Cistinosis	Medicamento: mercaptamina	Fase II

MedlinePlus en español temas relacionados: Trastornos genéticos, Nefropatías, Trastornos metabólicos

Tipo de estudio: De intervención

Diseño del estudio: Tratamiento, Inocuidad/Eficacia

Detalles adicionales del estudio:

Fecha de inicio del estudio: 3 de abril de 1986

Matrícula total esperada: 225

Requisitos para participar

Sexo: Ambos

Centros de estudio y contactos

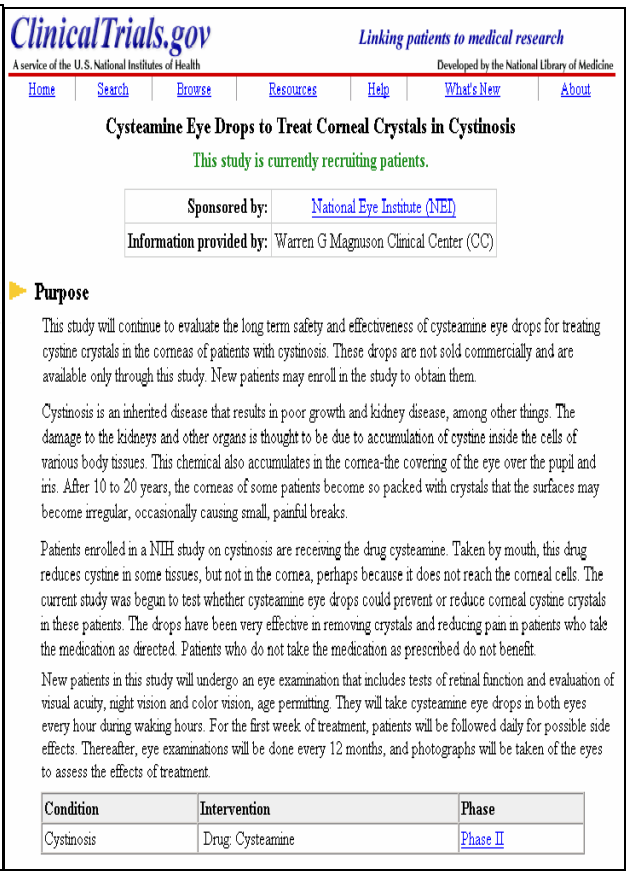
Maryland

National Eye Institute (NEI), 9000 Rockville Pike, Bethesda, Maryland, 20892, United States; En proceso de inscripción

Patient Recruitment and Public Liaison Office 1-800-411-1222

prpl@mail.cc.nih.gov

Figure 3.1 Spanish prototype doclet



ClinicalTrials.gov — Linking patients to medical research

A service of the U.S. National Institutes of Health — Developed by the National Library of Medicine

Home | Search | Browse | Resources | Help | What's New | About

Cysteamine Eye Drops to Treat Corneal Crystals in Cystinosis

This study is currently recruiting patients.

Sponsored by:	National Eye Institute (NEI)
Information provided by:	Warren G Magnuson Clinical Center (CC)

Purpose

This study will continue to evaluate the long term safety and effectiveness of cysteamine eye drops for treating cystine crystals in the corneas of patients with cystinosis. These drops are not sold commercially and are available only through this study. New patients may enroll in the study to obtain them.

Cystinosis is an inherited disease that results in poor growth and kidney disease, among other things. The damage to the kidneys and other organs is thought to be due to accumulation of cystine inside the cells of various body tissues. This chemical also accumulates in the cornea-the covering of the eye over the pupil and iris. After 10 to 20 years, the corneas of some patients become so packed with crystals that the surfaces may become irregular, occasionally causing small, painful breaks.

Patients enrolled in a NIH study on cystinosis are receiving the drug cysteamine. Taken by mouth, this drug reduces cystine in some tissues, but not in the cornea, perhaps because it does not reach the corneal cells. The current study was begun to test whether cysteamine eye drops could prevent or reduce corneal cystine crystals in these patients. The drops have been very effective in removing crystals and reducing pain in patients who take the medication as directed. Patients who do not take the medication as prescribed do not benefit.

New patients in this study will undergo an eye examination that includes tests of retinal function and evaluation of visual acuity, night vision and color vision, age permitting. They will take cysteamine eye drops in both eyes every hour during waking hours. For the first week of treatment, patients will be followed daily for possible side effects. Thereafter, eye examinations will be done every 12 months, and photographs will be taken of the eyes to assess the effects of treatment.

Condition	Intervention	Phase
Cystinosis	Drug: Cysteamine	Phase II

Figure 3.2 Corresponding English document

PRELIMINARY EVALUATIONS

An initial, informal review of the prototype by six bilingual medical translators and consumer health content providers elicited very positive reactions. This evaluation focused on linguistic issues as well as functional ones, such as results and performance. Some evaluators expressed a desire for more information in Spanish about the purpose of each trial ("Purpose" heading), but the consensus was that the abbreviated Spanish record is sufficient to convey the gist of the study. The evaluators indicated the functional advantage of a language toggle button on each Spanish and English page, already implemented in the Spanish site.

Subsequently, 96 bilingual speakers —members of the Center for Disease Control and Prevention (CDC) en Español Team, Association of Latino Employees at CDC (ATSDR), and the CDCSPAN Listserv — participated in a preliminary evaluation of the prototype. They were provided with guideline questions, as described in [11], with two additional topics: translation quality and perceived usefulness of the doclets. The respondents tended to be highly educated professionals, such as physicians, and 80% were from outside the U.S. Substantive feedback returned by 87 participants (as free text) was coded into five categories (Table 1).

Category	Sample Feedback
Search	• Errors/missing terms in BTL noted
Navigation	• Link logo to Home page suggested
Design	• Not mixing Spanish **and** English on the same page suggested • Spanish search details requested
Translation	• Inconsistencies reported • Differences between European and Latin American Spanish noted by participants in Spain
Other	• Spanish-speaking contacts in study centers suggested • Links to glossary of clinical trial-specific terms in Spanish requested

Table 1. Sample feedback by category

Regarding the abbreviated nature of the Spanish records, 24% felt that additional information should be translated (i.e., Purpose and/or Eligibility) and 35% thought that doclets are suitable only for professionals and others already used to reading in English. Overall, 40% considered that the abbreviated Spanish record is sufficient to convey the gist of the study.

While many of the comments were supportive of the CLIR approach, others provided thoughtful and relevant feedback. In addition to correcting the reported errors, adding important missing terms, and handling inconsistencies, we are considering numerous suggestions, some of which require further analysis, research, and evaluation.

CURRENT DEVELOPMENTS

Based on above mentioned preliminary feedback, we believe that the combination of cross-language information retrieval and Spanish doclets provides sufficient data for those native Spanish speakers with *some* knowledge of English, to obtain the gist of the information and determine relevance with respect to their needs. We have recently started the process of conducting a feasibility study with Spanish-speaking consumers to determine (1) the ability of the doclets to convey the gist of a study; (2) user-rated overall quality of the CLIR-based retrieval set of records; and (3) perceived experience of the Spanish ClinicalTrials.gov prototype site. Through user testing with Spanish-speaking health consumers, we expect to obtain data regarding the target audience's information needs, their ability to obtain relevant information easily and accurately from the doclets (compared to the full documents in English), and their satisfaction with the doclets. We intend to use the feedback to ensure that further development of doclets account for usability and accessibility issues identified by the intended users.

One enhancement to the doclets that we are currently considering involves the addition of a brief purpose statement (no more than three sentences) under the "Purpose" heading, while maintaining a link in Spanish to the full English document. The purpose statement, which will be extracted from the Brief Description section under "Purpose," will focus on the specific scientific goals or objectives of the study. Typically, these sentences are easily identified as many are introduced by a limited number of linguistic expressions (e.g., "this study will"). The identification of such linguistic patterns will facilitate greatly the task of developing natural language processing techniques to programmatically extract the key statements in the Brief Summary field. Based on experience with over 1,000 purpose statements, the present algorithm for extracting purpose statements has an accuracy rate of 92%. Once extracted, the purpose statements are manually translated into Spanish and will be included in the doclet.

The purpose statements represent the main point or "crux" of the study[7]. By "reducing" the Brief Description to an average of about 40 words per purpose statement, we bring this information closer to what is feasible for human translation, given the availability and extent of our resources. The same mechanism that was implemented for editing and translating the English titles into Spanish will have to be extended to include the purpose statements as well. To better assess the impact of adding the purpose statement to facilitate relevance ranking by the Spanish speaking users of ClinicalTrials.gov, the feasibility study will include doclets containing translated purpose statements.

The output of the algorithm using the four paragraphs of the Brief Summary from the document in Fig. 3.2 is as follows:

> This study will continue to evaluate the long term safety and effectiveness of cysteamine eye drops for treating cystine crystals in the corneas of patients with cystinosis. These drops are not sold commercially and are available only through this study. New patients may enroll in the study to obtain them.

The following figure will display the doclet on Fig. 3.1 but it will include the purpose statement (shown in English above) already translated into Spanish:

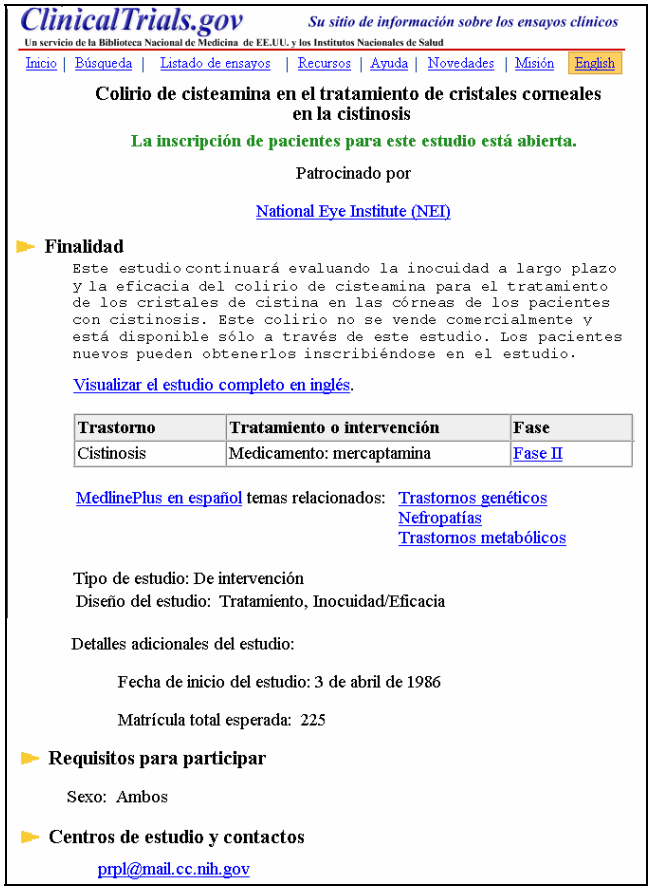


Figure 4: Spanish prototype doclet with purpose statement

CONCLUSION

To accommodate the needs of non-English-speaking users, we have begun to use CLIR to bridge the language gap for consumer health information systems. Because CLIR only addresses the retrieval of relevant documents in English, we have also developed display mechanisms to assist non-English

speaking users in accessing, understanding, and applying the health information they find. This case study describes one approach to providing Spanish-language support for ClinicalTrials.gov, through the integration of additional components into a system originally designed and implemented in English only. If further user tests support the usefulness and effectiveness of this CLIR approach, other consumer health information systems may consider adopting CLIR, as implemented here, or new and improved techniques to support non-English speakers.

ACKNOWLEDGMENTS

We thank Drs. Alexa T. McCray and Aaron B. Navarro for their insights, José A. Mendizabal, Multilingual Services, CDC for coordinating the feedback on the prototype, and Laurel Graham for many valuable contributions. This research was supported by the Intramural Research Program of the NIH, National Library of Medicine/Lister Hill National Center for Biomedical Communications.

BIBLIOGRAPHIC REFERENCES

[1] Berland GK, Elliott MN, Morales LS, et al. Health information on the Internet: accessibility, quality, and readability in English and Spanish. JAMA. 2001;285:2612-2621.

[2] Zeng Q, Kogan S, Ash N, Greenes RA, Boxwala AA. Characteristics of consumer terminology for health information retrieval. Methods Inf Med. 2002;41(4):289-98.

[3] Katz DG, Dutcher GA, Toigo TA, Bates R, Temple F, Cadden CG. The AIDS Clinical Trials Information Service (ACTIS): a decade of providing clinical trials information. Public Health Rep. 2002 Mar-Apr;117(2):123-30.

[4] Voge S. NOAH—New York Online Access to Health: library collaboration for bilingual consumer health information on the Internet. Bull Med Libr Assoc. 1998 Jul;86(3):326-34.

[5] Boyer C, Baujard V, Griesser V, Scherrer JR. HONselect: a multilingual and intelligent search tool integrating heterogeneous web resources. Int J Med Inform. 2001 Dec;64(2-3):253-8.

[6] Nelson SJ, Schopen M, Schulman J-L, Arluk N. An interlingual database of MeSH translations. [Cited 2004 Dec 16]. Available from: http://www.nlm.nih.gov/mesh/intlmesh.html.

[7] Hutchins J. Towards a new vision for MT. MT Summit VIII. 2001 Sep 18-22; Santiago de Compostela, Galicia, Spain. p. 1-6.

[8] Rosemblat G, Tse T, Gemoets D. Adapting a monolingual consumer health system for Spanish cross-language information retrieval. Proceedings of the 8th International ISKO Conference. 2004 Jul 13-16; London, England. p. 315-321.

[9] Rosemblat G, Gemoets D, Browne AC, Tse T. Machine translation-supported cross language information retrieval for a consumer health resource. Proc AMIA Symp. 2003 Nov 8-12; Washington, DC, USA. p 564-8.

[10] McCray AT, Ide NC, Loane RR, Tse T. Strategies for supporting consumer health information seeking. Medinfo. 2004 Sept 7-11; California, USA. p 1152-6.

[11] McCray AT, Dorfman E, Ripple A, Ide NC, Jha M, Katz DG, Loane RF, Tse T. Usability issues in developing a Web-based consumer health site. Proceedings AMIA Symposium 2000 Nov 4-8; Washington, DC, USA. p. 556-60.

[12] Logan R, Graham L, Rosemblat G, Tse T. Assessing how different factors affect text readability in the consumer health domain. In preparation.

---

[1] http://www.aidsinfo.nih.gov/
[2] http://www.noah.cuny.edu/
[3] http://medlineplus.gov/spanish/
[4] http://www.hon.ch/HONselect/
[5] http://clinicaltrials.gov/

[6]Some of the sources that we used included the Medical Subject Headings or MeSH (Bireme/U.S. National Library of Medicine), NLM Unified Medical Language System® (UMLS®), Freelang [database on the Internet] (availabe from: http://www.freelang.net/), clinical-trials vocabulary.
[7]Interestingly, the ability to communicate the main point appears to be one of the highest ranking factors that contributes to the ease of readability of consumer health documents [12]