

Increasing access to medical knowledge using multilingual search interfaces

By

Steven Bedrick

A DISSERTATION

Presented to the Department of
Medical Informatics & Clinical Epidemiology
and the Oregon Health & Science University

School of Medicine

in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

May 2011

School of Medicine
Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the PhD dissertation of

Steven Bedrick

“Increasing access to medical knowledge using multilingual search interfaces”

has been approved

Advisor: Dr. William Hersh, MD

Member: Dr. Nancy Carney, PhD

Member: Dr. Aaron Cohen, MD, MS

Member: Dr. Paul Gorman, MD

Member: Dr. Misha Pavel, PhD

Contents

Acknowledgments	x
Abstract	xiii
1 Introduction	1
2 Issues Facing NNES Researchers	6
3 Information Behavior of Clinicians	15
3.1 What kinds of information do clinicians need?	19
3.2 When, how, and where do clinicians get answers?	26
4 CLIR & Multilingual UI	36
4.1 CLIR Evaluation	44
4.2 CLIR in Biomedicine	59
5 Research Statement	67
6 The BuscTrad System	69
6.1 High-level architecture	70
6.1.1 Caching	74
6.2 Query	76
6.3 Search	77

6.4	Translation	79
6.4.1	Why Google Translate?	79
6.4.2	Implementation Details	81
6.5	Results Presentation	86
6.6	Pre-loaded result sets	89
6.7	Future Development	94
7	FREDO System	99
7.1	Evaluation plans	100
7.2	Collectors	104
7.3	System Flow	105
7.4	Surveys	112
7.4.1	Survey Syntax	114
7.4.2	Question Dependencies	120
7.4.3	Administration	121
7.5	Static Pages	124
7.6	External Instruments	129
7.6.1	Example external URL step: BuscTrad document selection . .	134
7.7	Future Development	136
8	Study Methodology	140
8.1	Survey instruments	142
8.1.1	Subject Demographic Questionnaire	143
8.1.2	Language assessment instruments	144
8.1.3	Post-task follow-up	149
8.2	Document Selection Task	150
8.2.1	Document collection	152
8.3	Hypotheses & Outcome Measures	156

8.4	Subjects & recruitment	161
8.5	Pilot Testing	163
9	Results	166
9.1	Subject Demographics	167
9.2	Language Assessment Results	179
9.3	Subject Preferences	187
9.4	Subject Performance	203
9.4.1	Classification Accuracy	203
9.4.2	Speed	212
10	Discussion	217
10.1	Preferences	217
10.2	Accuracy	221
10.2.1	Ground Truth Assessment	225
10.2.2	A Different Task?	234
10.3	Speed	245
10.4	Limitations	249
11	Conclusion & Future Work	256
A	Final FREDO evaluation plan	260
B	Survey Instruments: Demographics, English	263
C	Survey Instruments: Demographics, Spanish	271
D	ILR Self-Assessment Scale, English	279
E	ILR Self-Assessment Scale, Spanish	283
F	ILR Instrument Scoring Function	286

<i>CONTENTS</i>	iv
G Survey Instruments: Language, English	289
H Survey Instruments: Language, Spanish	303
I Survey Instruments: Post-task follow-up, English	317
J Survey Instruments: Post-task follow-up, Spanish	321
K Subject Recruitment Letter	325
Bibliography	327

List of Figures

3.1	The Leckie, Pettigrew, and Sylvain model of information behavior.	20
4.1	The BabelMesh query interface, including automatic query auto-completion.	61
4.2	The BabelMesh results presentation interface.	62
4.3	The Brazilian “Virtual Library of Health’s” Spanish-language search interface.	63
6.1	Major architectural modules of the BuscTrad system.	71
6.2	The flow of data within the search module (repository interface) of BuscTrad.	80
6.3	The public-facing Google Translate interface.	82
6.4	The flow of data within the translation module of BuscTrad.	87
6.5	Monolingual English results presentation module.	89
6.6	Monolingual Spanish results presentation module.	90
6.7	Bilingual results presentation module.	90
6.8	Translated MeSH heading display.	91
6.9	Query-term highlighting.	91
6.10	Pre-loading an article set.	94
6.11	A pre-loaded article set.	95
6.12	Viewing a pre-loaded article set from a study subject’s perspective.	96

6.13	Merging two existing article sets.	96
7.1	Setting up a new collector.	106
7.2	The collector administrative interface.	107
7.3	A subject participation log.	108
7.4	FREDO logs detailed technical information about every subject in- teraction.	109
7.5	The operational flow of a subject participating in an evaluation hosted using the FREDO system.	111
7.6	A simple survey.	113
7.7	A drop-down question interpolated within its label.	114
7.8	A set of answers generated by the <code>likert</code> macro.	119
7.9	An example of chained dependent questions.	123
7.10	The survey administration screen, showing the data dictionary. . . .	124
7.11	The survey administration screen, showing subjects' responses to the survey's questions.	125
7.12	The page entry and administration UI.	130
7.13	A simple page.	131
7.14	A more complex page.	132
7.15	The administrative screen for BuscTrad's document selection interface.	137
8.1	System evaluation protocol.	142
8.2	Document set creation.	154
8.3	Performance operator characteristic curve sets for three possible hy- pothetical strategy combinations.	161
8.4	Certificate of completion.	163
8.5	Pilot testing an early version of BuscTrad at Hospital Luis Vernaza in Guayaquil, Ecuador, November 2009.	165

9.1	QUORUM diagram showing subject inclusion flow.	168
9.2	When do subjects leave?	169
9.3	Subject age distribution.	170
9.4	Countries in which study participants reported primarily working. .	171
9.5	Reported expertise at diagnosing & treating traumatic brain injury, by medical role.	173
9.6	The responses of the nine subjects who skipped at least one of the ILR instrument's questions, along with their scores.	182
9.7	Distribution of ILR score among subjects.	184
9.8	Distribution of self-assessed proficiency for reading English in non- medical contexts.	184
9.9	Distribution of self-assessed reading proficiency for for reading En- glish in medical contexts.	185
9.10	Subject ILR scores vs. years of formal English education. Note the weak but significant correlation.	185
9.11	Data from Table 9.16, stratified by ILR score.	186
9.12	Subjects' unstratified responses to the usability feedback questions. .	188
9.13	Density plot of subject responses by ILR level.	198
9.14	Dichotomized subject responses to the usability feedback questions.	199
9.15	Proportion of subjects preferring the "English-only" interface to the other two choices for the five usability feedback questions, stratified by ILR.	200
9.16	Proportion of subjects preferring the "Spanish-only" interface to the other two choices for the five usability feedback questions, stratified by ILR.	201

9.17	Proportion of subjects preferring the “Bilingual” interface to the other two choices for the five usability feedback questions, stratified by ILR.	202
9.18	Histograms of the raw number of document selections made by subjects.	207
9.19	Subject selection counts from the English-only task plotted against counts from the Spanish-only and Bilingual tasks.	208
9.20	Distribution of precision scores across the three interface modes.	209
9.21	Distribution of recall scores across the three interface modes.	209
9.22	Distribution of F-measure scores across the three interface modes.	209
9.23	Within-subject F-measure correlation.	210
9.24	Mean F-measure for each task, stratified by ILR score.	211
9.25	Mean F-measure for each task, stratified by ILR score, including error bars representing 95% confidence intervals about each data point.	211
9.26	Histograms of elapsed time on the document selection task	216
10.1	False positive rates, by task.	222
10.2	Selection count, by article.	223
10.3	Selection count, by article, including relevance status.	225
10.4	Selection count, by article, for the three secondary raters.	230
10.5	Mean F-measure, by interface mode and ILR, using the secondary raters’ judgments.	233
10.6	Article selection count, by interface mode, color-coded according to the “Title contains TBI” rubric.	235
10.7	Mean F-measure, by ILR and interface mode, using the “Has TBI” computed relevance judgements. Compare to Figure 9.24.	244

10.8 Mean F-measure, by ILR and interface mode, using the “Has TBI”
computed relevance judgements and including error bars represent-
ing 95% confidence intervals. Compare to Figure 9.25. 245

Acknowledgments

I wish to offer my sincere and heartfelt thanks to my entire advisory and examination committee, all of whom played important roles in making this dissertation possible, beginning with Dr. William Hersh, my advisor. I thank Dr. Hersh for his constant support and guidance during the entire process, and would also like to acknowledge his vital role in enabling my Latin American travels, through which I learned first-hand about the information challenges facing Latin American clinicians. Thank you, Bill.

My other committee members were just as essential to my dissertation's success; in no particular order, I thank: Dr. Aaron Cohen for generously sharing his time and expertise, both academic and otherwise; Dr. Paul Gorman, for providing invaluable insights at all stages of the process and for serving as a sounding board for new ideas; and Dr. Misha Pavel, for possessing and sharing an astonishing depth and breadth of knowledge regarding experimental design. Last, but emphatically not least, I thank Dr. Nancy Carney for inspiring me to ask these questions in the first place, by taking me with her to Argentina in 2006; fittingly, she also provided essential guidance about how to tie the whole project together in the end.

In addition to my committee, I wish to acknowledge all of my study participants, who generously gave me more of their time than they had probably intended. I would also like to thank my Latin American collaborators, both for mo-

tivating this study and for making its completion possible. In particular, I thank: Dra. Paula Otero and the rest of the medical informatics department at Hospital Italiano in Buenos Aires, Argentina; Dr. Gustavo Petroni and the critical care team at Hospital de Emergencias Clemente Álvarez in Rosario, Argentina; and Dr. Alejandro Mauro at MEGASALUD, in Santiago Chile. Many of my subjects assisted with subject recruitment; I would like to specifically acknowledge Dr. Daniel Hinojosa and Gloria Niño Alape for their help in enrolling additional participants.

Dra. Vanina Taliercio, Dr. Damian Borbolla, and Amy Huddleston provided invaluable assistance in translating my study's materials, particularly the survey instruments, into Spanish. Natalia Pereira gave helpful feedback regarding the wording and formulation of Spanish-language survey questions.

This work was supported and made possible by a National Library of Medicine informatics fellowship (#T15LM007088), for which I am eternally grateful. While the financial support itself was necessary, welcome, and appreciated, the true value of the fellowship lay in the fact that it provided me with a superb group of colleagues with whom to spend five years of my life. I thank all of the OHSU DMICE fellows, past and present, for their friendship and support.

I would also like to thank the entire faculty and staff of the Department of Medical Informatics & Clinical Epidemiology, all of whom have contributed to an exceptionally collegial and stimulating academic environment. In particular, I wish to acknowledge Dr. Jayashree Kalpathy-Cramer, whose friendship, encouragement, and advice have been instrumental to my development as a scientist and as a person, and whose analytical skills and patience have helped me through many challenging statistical situations.

Finally, I would like to thank my family and friends, for sticking by me in my pursuit of a doctorate and for stoically enduring my frequent absences, both physical and mental. And, of course, special thanks go to my partner, Candice Erdmann,

whose unwavering support formed the bedrock upon which this dissertation was built.

Abstract

Clinicians have numerous and diverse information needs, and face a similarly-diverse variety of obstacles preventing those needs from being met. Some of these obstacles are technical in nature, while others are organizational or educational. For many of the world's clinicians, however, one of the most important obstacles is linguistic: the vast majority of internationally-published medical and scientific literature is written in the English language, which means that many potential users of this content are unable to do so without expending the significant additional time and effort required to read a foreign language. Machine translation and cross-language information retrieval may be able to assist non-native-English-speaking clinicians who wish to make use of English-language medical literature; however, there is relatively little research about how such technologies might be adapted for use in clinical settings, or how such adaptations might be evaluated.

This dissertation describes a novel bilingual search interface that uses machine translation to present English-language MEDLINE search results in Spanish, along with an evaluation protocol that seeks to assess the effects of such an interface on subjects' ability to identify relevant search results. The protocol involves survey instruments, an English proficiency assessment, and a series of simulated document retrieval tasks that take place under several different user interface conditions (monolingual-English, monolingual-Spanish, and bilingual). The protocol's outcomes of interest include both subject preference as well as subject performance.

We also describe a novel computer system designed to assist investigators conducting large-scale data collection activities over the Internet.

We conducted a pilot evaluation of our search system using this experimental protocol, in which approximately 60 Latin American clinicians participated. We found considerable diversity of opinion amongst our subjects: some subjects expressed a preference for the search interface with no translated content, while others strongly preferred a bilingual interface in which translated results were shown alongside untranslated (i.e., English) equivalents. A smaller, yet sizable, segment of our subjects preferred a system that presented translated search results without any English content at all. Subjects' preferences were strongly related to their English reading proficiency; subjects with weaker English reading skills were much more likely to prefer the monolingual-Spanish interface than subjects with stronger English proficiency, and, inversely, subjects with stronger English reading skills were more likely to prefer the monolingual-English interface. Subjects who preferred the Bilingual interface tended somewhat to have low-to-medium English reading skills; however, there was a sizable contingent of higher-proficiency subjects who preferred the Bilingual mode over the monolingual-English mode.

We observed differences between interfaces in terms of subject performance at the document selection task, but methodological challenges prevented us from characterizing or fully understanding the nature of those differences. Our experience highlights the difficulties inherent to any study involving real-world users, and illuminates several specific challenges facing information retrieval studies that cross cultural and linguistic boundaries. With some modifications, our protocol and survey instruments may be used to investigate a wide variety of questions regarding cross-language information retrieval user interfaces.

Chapter 1

Introduction

The practice of medicine is heavily dependent on access to information. Clinicians of all stripes are constantly seeking out information, whether in response to a specific clinical information need, for self- or patient-education, or for other reasons altogether. There are many different types of resources to which clinicians may turn to address their information needs: their colleagues, text-books, electronic databases, and so on[1, 2]. One of the most important information sources, especially from the perspective of evidence-based medicine[3], is the body of work that makes up the “medical literature:” published, peer-reviewed accounts of research, literature reviews, metaanalyses, and the like.

It is well known that actual, practicing clinicians often face difficulties in making use of primary-source medical literature (see chapter 3 for a more detailed discussion on this theme). These difficulties can arise from a number of sources: clearly framing a clinical question can be a difficult task for clinicians[4, 5], and so is the process of actually using a search system to find and retrieve answers to those questions[6, 7]. However, many clinicians face an even more fundamental set of obstacles when attempting to make use of medical literature: language.

The majority of medical and scientific articles are published in English[8], and

while many non-native-english-speaking (NNES)¹ clinicians are proficient to some degree at reading in English, individual levels of proficiency vary greatly from person to person and country to country. Furthermore, the user interfaces of most popular literature search engines are English-only. We therefore believe that language represents a significant barrier for NNES clinicians who wish to make use of the international medical literature.

This dissertation is about ways that machine translation technology may be used to help NNES clinical users of literature search systems overcome this language barrier, and make better use of published articles, guidelines, and other such content. Its focus is specifically on Spanish-speaking clinicians, for several reasons. The first and foremost reason is the fact that Spanish is a regionally important language: Outside of the United States and Canada, Spanish is the dominant language of the Western Hemisphere, and is spoken by more than 300,000,000 people[9]. Within the United States and Canada, Spanish is quickly growing in prominence, and is spoken at home by more than 35,000,000 people[10]. As researchers who have been privileged to speak English as his native language, we find the improvement of professional communication with our NNES colleagues to be a worthy goal in and of itself— to say nothing of the potential benefits for clinical practice and research in Spanish-speaking parts of the Americas resulting from increased access to medical literature.

A second reason for focusing on Spanish-speaking clinicians is technical in nature. Machine translation between English and Spanish is relatively mature, and has been found to perform better than between other language pairs[11]. The two languages, while quite different, are more closely related in linguistic terms than, say, English and Japanese, or English and Arabic, and share certain other properties that make it an ideal first step in exploring the possible uses and effects of

¹ A non-native English speaker is an English-speaking individual whose native language is not English.

automated translation support on user behavior.

A final reason for focusing this dissertation's work on Spanish-speaking clinicians is that of convenience. The Department of Medical Informatics and Clinical Epidemiology has several notable and close working relationships with Spanish-speaking medical organizations, both in terms of our research mission as well as our educational program. Given this history of collaboration with Latin American colleagues, it is reasonable for us to begin our research in this area by focusing on the population with which we are already working. Besides these relatively abstract benefits, our existing collaborations have a practical benefit, as well: our pool of colleagues has proven to be an invaluable source of beta testers and experimental subjects.

There have been successful attempts to build multilingual search interfaces to the biomedical literature[12, 13, etc.], and there also exist several Spanish-language biomedical literature search engines². Both approaches, however, deliver their results in the original language of publication (i.e., English), and therefore do little to support users with lower levels of English reading proficiency.

To address this problem, many users rely on linguistic support tools (dictionaries, machine translation systems, etc.) that are not well-integrated into the search process and are of variable quality. Furthermore, very few studies have evaluated the relative efficacy of native-language vs. foreign-language search *interfaces* (see Section 4.1), and none have looked at bilingual or machine-translated results presentation in a medical setting. Therefore, there exists a clear need both for a more integrated and linguistically supportive literature search tool as well as for additional evaluation of user performance and behavior under multilingual conditions.

The goal of this dissertation is therefore to explore the effects of multi- and native-language interfaces on the results analysis step in the search process of na-

²The Brazilian Biblioteca Virtual de Saúde, for example, offers Spanish and Portuguese interfaces to MEDLINE[14].

tive Spanish-speaking clinicians. Our overarching hypothesis is that *non-native-English-speaking users will perform better in interface modes containing native-language text than they will in monolingual English-only modes*. In order to investigate this hypothesis, we built an experimental, publicly-accessible multilingual information retrieval system (named *BuscTrad*³), and carried out a user study that evaluated the effects of different user interfaces on subjects' ability to identify relevant articles from a set of simulated search results. Along the way, we also developed (and used) a novel and highly flexible software tool for managing large-scale web-based user studies. This dissertation's structure is as follows:

Chapter 2 discusses some of the issues facing NNES clinicians and researchers, and gives a brief overview of the literature on the subject.

Chapter 3 reviews the subject of clinical information behavior, and includes discussions of what information clinicians need; where, when, and how they satisfy those needs; and explores the role that published primary-source medical literature plays in that process.

Chapter 4 discusses several of the most important approaches that have been developed to support second-language users of information systems, along with discussion of various ways in which they have succeeded or failed. It also discusses cross language information retrieval (CLIR), with a particular emphasis on CLIR in biomedicine.

Chapter 6 describes *BuscTrad*, the aforementioned multilingual multilingual information retrieval system.

Chapter 7 describes the novel user study management system developed as part of this work.

Chapter 8 details the experimental protocol behind the aforementioned user study, while Chapter 9 presents the results from the same.

³A portmanteau of the Spanish verbs "buscar" ("search") and "traducir" ("translate").

In Chapter 10, we discuss those results as well as the study's limitations, and Chapter 11 contains a summarization and discussion of future work.

Chapter 2

Issues Facing

Non-Native-English-Speaking

Researchers and Clinicians

To quote Ulrich Ammon, the observation that English “is today’s dominant language of science is stating what would be called a *Binsenweisheit* in German, a trivially obvious insight.”[15]. Throughout history, it has been common for one language or another to dominate scientific discourse; various languages have flourished in their season, and today, for a variety of reasons[16], it is English’s turn.¹ This is particularly true in biomedicine, as the vast majority of internationally-published biomedical research is in the English language, and the relative proportion of English-language articles has been increasing over time. Loria and Arroyo’s survey of MEDLINE entries found that, from 1966 to 2000, the percentage of MEDLINE-indexed articles written in English increased from 50% to 90%[8]. Monge-Nájera and Nielsen found similar rates of change in their survey, as well[18]. Loria and Arroyo further found that an increasing number of journals from non-

¹ Had it not been for World War II, today’s scientific *lingua franca* could well have been German[17]!

Anglophone countries are publishing English-language articles.

However, although non-English-language articles represent a minority of MEDLINE's content, it is an important minority. Moher, et al. found that non-English-language randomized controlled trials (RCTs) were just as methodologically sound as English-language RCTs[19], and in a later study found that excluding non-English RCTs from systematic reviews and meta-analyses could significantly impact those studies' conclusions[20]. In 2002, Clark and Castro examined 62 systematic reviews published in five high-impact English-language journals in 1997, and found that in 71% of the reviews, the original authors had missed potentially relevant articles by failing to search the *Literatura Latino Americana e do Caribe em Ciências da Saúde* (LILACS) database, which indexes articles from regional journals in Latin America and the Caribbean.[21] Furthermore, in a 2008 editorial, Isaac Fung suggested that systematic reviews in epidemiology— which often cover topics of a particularly global nature— ought to explicitly include articles from as many languages as possible.[22] Obviously, the issue of English's dominance in scientific discourse is global in nature. However, for the purposes of this chapter, we will focus primarily on Latin America and the Caribbean.

There has been a great deal of research conducted investigating the reasons that non-native English-speaking (NNES) researchers choose to publish in English rather than in their native languages. Essentially, it comes down to the fact that articles written in English will automatically have a much larger potential audience than articles written in another language. The editors of the *Brazilian Archives of Ophthalmology* (which saw an increased number of English-language submissions after being picked up by MEDLINE) sees the matter as one of authors wishing to maximize their international exposure[23], and they are far from alone in this viewpoint. Authors in non-Anglophone countries are no less sensitive to the demands of impact factor and other such metrics than are Anglophone authors, and articles

in local or regional (i.e., non-international) journals are far less likely to be cited than articles in major journals, as are articles in those journals that are not written in English.[24] This fact has been shown to affect the publication choices of some researchers; Egger, et al.[25] showed that German-speaking authors were more likely to publish the results of randomized controlled trials in English-language journals when their results were significant, and would publish their negative or null results in German-language journals.² Meneghini and Packer summarize the overall dilemma facing NNES scientists thusly:

Scientists look for international visibility by publishing in English either in national journals or in high-profile international journals; conversely, they hope to attract a larger regional audience by publishing in their mother tongue or they choose a national journal because they are not sufficiently fluent in English.[26]

In addition to author-oriented reasons for publishing in English, however, some researchers have identified editorial pressure on NNES authors to publish in English. Loria notes that one of the few remaining Mexican scientific journals recently converted to an English-only format[8], and in the same article note that the Mexican National Council of Research and Technology (which is roughly analogous to the United States' National Science Foundation) "specifically asks for the exclusion of any information from papers published in local journals" from its researchers' annual status reports, "which is the same as telling researchers to publish only in English and in foreign, developed countries."

Of course, when NNES authors do attempt to publish in English-language journals, they face an uphill battle. Some of this is due to the difficulties inherent to writing in a foreign language; indeed, several studies have shown that a country's

² The authors noted that this publication pattern could seriously affect the results of any systematic reviews or meta-analyses that only included English-language articles.

overall level of English proficiency is strongly related to how well it is represented in the international scientific literature[27, 28]. However, some of the problem may result from prejudice and/or bias on the part of reviewers. A 1998 study by Link[29] found that US-based editorial reviewers tended to rank papers from US-based authors much more highly than papers from non-US authors. Several studies have investigated this phenomenon, and have generally concluded that while some bias is certainly present, a large part of the problem can actually be attributed to the general quality of the writing (as opposed to the quality of research) in many submissions from NNES researchers[30, 31]. One editorial by the editor of *EMBO Reports* also pointed out that NNES authors are at a disadvantage when it comes to writing the cover letters that often accompany manuscripts, and as such may not be as readily able to “grab the attention” of editors.[32] Additionally, some of the editorial bias may be due more to subject matter than to national or linguistic origin: one study that talked to NNES researchers directly found that many authors felt that Anglophone journals were less interested in publishing articles about medical conditions that are prevalent primarily in less-developed settings (Typhoid, Dengue Fever, etc.), and quoted one researcher as saying that Anglophone journals were “heavily biased towards high-tech lab-oriented medical research.”[33] However, in that same paper, Horton quotes several researchers who saw their acceptance rates plummet when they moved from the US back to their home countries. Clearly, editorial bias is a complex and multifaceted problem.

Finally, there is even some evidence to suggest that, once NNES researchers *do* get published in a mainstream, English-language journal, they still face difficulties in having their work widely read. Several authors ([34, 35]) have noted that MEDLINE and other such databases often have difficulty properly indexing non-English names, making retrieval difficult and inconsistent, and at least

one study has found significant problems with the English translations of the titles of Spanish-language articles as indexed in MEDLINE[36]. While this is certainly a problem, however, it is a relatively benign one compared to the findings of Meneghini, et al.[37], who conducted a survey of articles from several prestigious journals, and compared the impact factors (IF) of articles from four Latin American countries against articles from five “developed countries.” Meneghini found that the articles from Latin American authors had significantly lower impact factors than did the articles from the developed subgroup. The Latin American articles’ IF scores averaged 66% of that of the journal as a whole, whereas the IF scores for the articles from the “developed countries” closely tracked that of the journal. Furthermore, Meneghini found that this effect appeared to be related to whether or not the Latin Americans collaborated with researchers from outside Latin America: collaborative articles showed no change in IF compared to articles from the developed countries, whereas non-collaborative articles (i.e., articles authored entirely by Latin American authors) had significantly smaller IF scores.

This overall linguistic pattern, then, effectively limits participation in scientific research to individuals with particularly strong English abilities,³ and results in a diminished scientific output from non-Anglophone countries (and a corresponding underrepresentation of NNES researchers’ work).[39, 27, 31, 8, 18] Various ideas have been put forth about how to improve NNES researchers’ ability to be published in English-language journals. Perhaps the most extreme idea is that of Charlton, who suggests that authors be allowed to “patch together” large sections of scientific articles by directly quoting (with proper attribution, of course) existing articles “whenever the author judges that these quotations are a precise and clear exposition of what s/he would like to say – if only they had the linguistic competence.”[40] While this idea no doubt has some appeal to anybody who has

³Albert[38] points out that the burden of writing in English is sufficient to discourage many NNES researchers—specifically clinicians—from publishing at all!

struggled to assemble a literature review, it is problematic both in that it represents a major shift from standard editorial practice as well as that it only solves part of the problem. Yes, it would make it somewhat easier for NNES authors to write the “background” sections of their papers, but this is arguably the *least* important part of a scientific article. Authors would still have to face the challenging task of summarizing their research and discussing its implications in a foreign language, and the uneven, patchwork nature of such an article would undoubtedly have a significant effect on its readability— which would do little to address the previously-discussed editorial challenges facing NNES-authored articles.

More realistic suggestions have largely centered around providing additional writing assistance for NNES authors, typically in the form of either workshops[41] or outright editorial mentorship[42]. These have the potential to be very useful for NNES authors; Vasconcelos, et al.[28] point out that in many developing countries, coursework in professional or scientific writing is not a part of many academic programs training clinicians and researchers, and suggest that greater attention be paid by non-Anglophone countries’ educational systems to this topic. Many scientific journals have also published articles offering advice on how best to write scientific English, aimed primarily (but not exclusively!) at NNES authors; those by Lanza[43], Kirkman[44], and Tompson[45] represent three excellent examples of this.

Our discussion to this point has focused on some of the problems that the dominance of English causes NNES researchers who wish to publish their research and participate in the international discourse of science. We have not addressed the much larger issue of NNES *consumers* of scientific knowledge, particularly those in a clinical setting. The body of literature studying the effects of the language barrier on literature consumers is far smaller than that studying the effects on literature producers; However, there is growing recognition of the fact that, as more and

more NNES authors publish in English, their co-linguists who are less proficient in English are being “left behind.”[26] The body of literature discussing English proficiency levels among NNES researchers is extremely limited; one 2008 study by Vasconcelos, et al. found that out of more than 52,000 Brazilian scientists, only one third felt themselves to be “fully proficient” at English (meaning that they rate their English reading, writing, speaking and listening abilities as “good” or better).

In 2008, Isaac Fung suggested that technology and newer publishing practices (such as “open access” publication and wikis) may help to address this issue.[46] Along with several other ideas, he proposes that journals recruit “translator-editors” from among their readership much in the same way that they currently recruit peer reviewers, and that these individuals could be put to work translating abstracts and titles into a wide variety of languages. He also suggests that authors be encouraged by journals to include versions of their articles in both English as well as their mother tongues,⁴ and that both versions be made available to readers. This approach is also popular in many Spanish-language publications[47], as well as with any publications sponsored by the Pan-American Health Organization (PAHO), and some Canadian journals publish both English and French versions of their titles, abstracts, and/or articles.

The fact that these examples are exceptions to an overwhelming norm underscores the fact that language represents a significant barrier to NNES researchers and clinicians wishing to read the bulk of the published literature. Before individuals encounter this barrier, however, clinicians must first traverse a different barrier: they must *find* the articles. As previously discussed, MEDLINE’s content is primarily in English, and it must be noted that while MEDLINE entries do record the original language of each article, each entry’s title and abstract are indexed in English. As such, users querying MEDLINE— even if they are specifically searching

⁴ Since 2006, this has been the standard practice among the *Public Library of Science* family of journals.

for non-English content— must build their queries in English. There are alternative information sources, however. The previously-mentioned LILACS database, managed by the the Latin American and Caribbean Center on Health Sciences (also referred to by its Brazilian acronym, BIREME) indexes a very large number of articles in both Spanish and Portuguese, both from regional as well as international journals, and it provides Spanish, Portuguese, and English versions of its search interface. As excellent a resource as LILACS is, however, it is not as widely utilized as it perhaps should be— among the subjects we enrolled for the study described by this dissertation, only one or two reported regularly using LILACS (see Section 9.1), a finding consistent with that of Ospina, et al. who found that only 6% of 185 Latin American biomedical researchers routinely used LILACS.[48] Chapter 4 discusses technological solutions to the problem of searching for information across language barriers.

This chapter has demonstrated that language does indeed represent a major challenge facing NNES researchers and clinicians. However, we would like to close by pointing out that there exists another barrier that is in some ways even larger: access to the content itself. Subscriptions to scientific and medical journals can be extremely expensive, and while many publishers have different pricing scales for institutions in “developing” countries, the costs can still be prohibitive. The World Health Organization’s HINARI (“Programme for Access to Health Research”) program is a collaboration between more than 150 publishers and the WHO that aims to make the full-text content of more than 7,500 journals and other resources available to “local, not-for-profit institutions in developing countries.”[49] Another such program is BIREME’s SciELO (Scientific Electronic Library Online).[50] Initially begun as a database of full-text content from Brazilian journals, it has spread to include full-text content from regional journals from a number of other Latin American and Caribbean countries, and also includes sig-

nificant amounts of content from MEDLINE-indexed journals as well. Journals indexed by SciELO report increased numbers of submissions[51]⁵, and although its search engine is not without limitations (as discussed by Curioso, et al.[52]) it remains an important resource to researchers both in Latin America as well as in the rest of the world.[26]

While HINARI and SciELO do important work and are certainly both worthwhile programs, we can report anecdotally that relatively few Latin American clinicians of our acquaintance know of their existence. In fact, of the Latin American clinicians with whom we are personally acquainted, virtually all have expressed frustration at how difficult it can be to access the full-text of medical articles, and 87% of the subjects in Ospina's study reported that they had not included important references in their published articles because of problems obtaining the full text of the articles. Clearly, there remains much work to do in terms of increasing NNES scientists' and clinicians' ability to fully utilize internationally-published scientific and medical literature, and helping to lower the language barrier can only solve part of the problem.

⁵ In that article, Blank, et al. note that while being indexed in SciELO resulted in more Brazilian submissions to *Jornal de Pediatria*, being indexed by MEDLINE resulted in an increase in both Brazilian and foreign submissions.

Chapter 3

Information Behavior of Clinicians

The practice of medicine is highly dependent on timely and relevant information. However, given the infinite variety of patient histories and presenting problems, as well as the world's constantly expanding body of scientific knowledge, it is impossible for any single individual to know all the necessary information for every encounter. There is therefore often a gap between what, at any given point in time, a clinician knows and what he or she needs to know. Just how large is this "information gap"? As with many questions, the answer to this question depends on what we consider to be its "denominator": how much information must a clinician be familiar with to perform his or her work?

In 1979, Hodgkin identified a "diagnostic vocabulary" of nearly 500 concepts used by a sample of physicians[53]. While 500 concepts may sound like a lot for an individual to keep track of, Hodgkin's vocabulary was in fact far from complete, even by the standards of medicine in 1979. He was not attempting to proscriptively generate an exhaustive list of clinical concepts that clinicians needed to be aware of; rather, he was simply attempting to describe the "every-day" sorts of topics that a working physician would encounter and need to know how to deal with during a "normal day."

In 1976, Pauker et al. published[54] their oft-cited (cf. [55, 56, 57], etc.) finding that a complete understanding of internal medicine required some 2,000,000 unique facts.¹ Given the significant advances in medical knowledge that have taken place during the last thirty years,² this number can only have increased.

Thus, we have a very wide range in estimates as to the size of the information problem in medicine.

As such, clinicians regularly must seek out additional information as they carry out their work. The questions of when, why, and how they do so are all facets of the larger topic of *clinical information-seeking*, which continues to be an active field of study. This chapter will give an overview of existing research into and models of clinical information behavior, with a particular focus on how clinicians interact with electronic information sources.

We will begin with some vocabulary. Case defines “information seeking” as “a conscious effort to acquire information in response to a need or gap” in one’s knowledge, and categorizes it under the larger subject of “information behavior,” which is defined as “Information seeking, as well as the totality of other *unintentional* or *passive* behaviors... as well as purposive behaviors that do not involve seeking, such as actively *avoiding* information.”[59]

Another important concept to define clearly is that of the “information need.” There are many ways to think about what constitutes an information need (Case[60] identifies four major families of definition). Two models of information need that are particularly relevant to clinical settings are Belkin’s “Anomalous states of knowledge” [61] and Dervin’s “gaps.”[62] Under Belkin’s model, an information need is

¹ In that same paper, Pauker asks the question of whether 2,000,000 facts could be stored “at a reasonable cost” in a computer’s memory. After calculating that each fact would require ten words of computer memory, he concludes that it would be possible, if one were using “an inexpensive mass storage device (such as a magnetic disc or drum),” and that “even at present day prices, the cost would probably be no more than \$20,000”.

²Wyatt[57] quotes Ziman[58] as calculating the “doubling time” of medical information as being 19 years.

a cognitive concept, occurring when an individual recognizes that their “conceptual state of knowledge... is anomalous with respect to some goal.” By “anomaly,” Belkin means that the individual’s state of inadequacy with respect to their goal “could be due not only to lack of knowledge, but many other problems, such as uncertainty as to which of several potentially relevant concepts holds in some situation.”[61].

Dervin’s understanding of what constitutes an information need is definitely related to Belkin’s, but is more general and, for lack of a better term, “holistic.” Dervin’s general theory of information behavior is known as “Sense-Making,” and includes not just an individual and his or her task, but also includes the *context* surrounding both the individual and the task.[62] Dervin’s model puts the user of the information at the center, and frames their quest for information as part of a larger process of making sense of their world. She describes an “information need” as follows: An information need, she says,

implies a state that arises within a person, suggesting some kind of gap that requires filling. When applied to the word information, as in information need, what is suggested is a gap that can be filled by something that the needing person calls “information.”

[63] as cited in [59]

Note that Dervin’s definition makes no assumptions about what *kind* of information ultimately fills that gap, or how thoroughly the gap is filled; Dervin’s model is extremely user-focused, and if in the end the user considers their need to be satisfied, the model considers the matter resolved.

Many researchers have formulated theoretical models as to how humans go about resolving information needs. While these models often differ in their details, they typically have certain elements in common. Often, they include some number of what Byström and Järvelin’s model[64] refer to as “Personal Factors:”

demographic attributes of the user that affect their interest in resolving— or their ability to resolve— their information needs. Some models (e.g., Wilson’s Second Model[65]) go into considerable detail, while others (e.g., Johnson’s model[66] or Byström and Järvelin’s model) do not.

One model particularly well-suited to the medical domain is that of Leckie, Pettigrew, and Sylvain[67], which is reproduced in Figure 3.1. Note that while this model does not devote large amounts of attention to the user’s demographic or psychological background, it does include discrete concepts of “Work Roles” and “Tasks.” This demonstrates its professional focus, and reflects the fact information professionals (such as clinicians) often have multifaceted jobs, and that each professional facet will involve a different set of information needs (and tools available with which to meet those needs).

Another useful feature of this model is that its iterative nature reflects reality in that it views information-seeking as a dynamic process. In this, it is particularly reminiscent of many information behavior models designed specifically to model the behavior of users of information-retrieval systems. One prominent example of such a model is Marchionini’s “Information-seeking Process”[68], which focuses primarily on the process of obtaining information using an electronic information system. Marchionini’s model divides the process into eight steps: recognizing and accepting that one faces an information need, defining the problem, selecting a source, formulating a query, executing that query, examining the results, extracting information, and reflecting on that information.

One particularly valuable aspect of Marchionini’s model is that it describes not only the steps themselves, but also the possible transitions between each step, which he groups into three categories: the “default” transitions, which run unidirectionally between the steps as described; “high probability” transitions, which connect steps that a user is likely to repeat frequently as their search continues (e.g.,

from “Examine Results” to “Formulate Query,” or from “Extract Information” back to “Define Problem”); and “low probability” transitions from one state to another, which represent valid but less likely transitions (e.g., from “Extract Information” back to “Select Source”).³

For our purposes, we are at this point most interested in the early stages of these models: recognizing information needs and formulating questions. The next section discusses what sorts of information needs clinicians routinely encounter during their day-to-day work.

3.1 What kinds of information do clinicians need?

Margaret Thompson asserts that physicians

seek information for two main reasons: (1) to obtain answers to patient-specific questions that cannot be answered through their personal knowledge alone and (2) to stay abreast of developments in clinical medicine.[72]⁴

While this formulation is admirably simple and to-the-point, it is somewhat reductionist in nature. In 2004, using a qualitative approach, Bryant described a somewhat richer hierarchy of “perceived information needs,” or “information wants,” from general practitioners: clinical care, keeping up-to-date, information for patients, pharmacological information, gaps in knowledge, curiosity, and uncertainty [74]. Without exception, each of her informants reported “clinical care”

³As influential as Marchionini’s model is, it is far from the only model of information-seeking; for further discussion of models of information-seeking, we suggest consulting either Case’s “Looking for Information”[69], Fisher’s “Theories of Information Behavior”[70], or Hearst’s discussion of the topic in “Search User Interfaces”[71].

⁴ These activities can be thought of as being roughly analogous to Krikelas’s concepts of “information seeking” and “information gathering”[73].

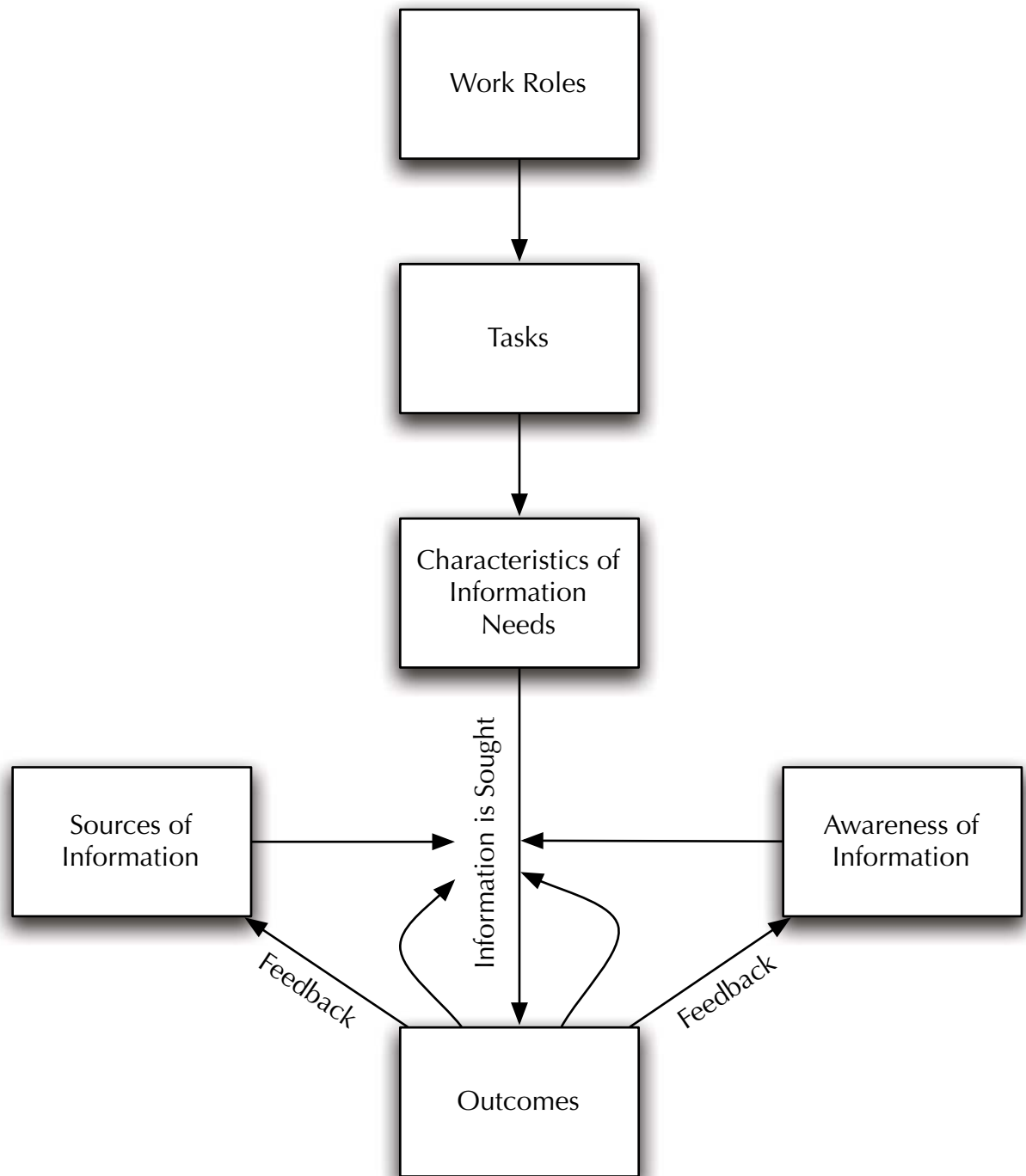


Figure 3.1: The Leckie, Pettigrew, and Sylvain model of information behavior, originally from [67] and reproduced from [66].

as being their primary reason for seeking information; there was more variation among the relative rankings of the other listed information needs.

The Bryant hierarchy is an improvement over Thompson's simplistic model, but it is still rather high-level. Many studies of clinician information needs have focused in greater detail on the specific information gaps (often referred to as "clinical questions") that arise during various clinical situations. Sometimes, those performing such studies are seeking to evaluate a tool or resource meant to help answer clinical questions⁵, and are less interested in the questions themselves than in the processes used to address them. For example, Chambliss & Conley[75] and Gorman et al.[76] both collected "*in vivo*" clinical questions by directly observing physicians as they did their daily work, but although both studies were certainly interested in documenting the questions, their primary goals were to explore the extent to which existing resources (MEDLINE, etc.) were able to address those questions.⁶

Other studies have been more descriptive in nature, and have sought to identify and systematize the sorts of questions that arise during clinical care. These studies typically attempt to measure or capture four different types of information: clinical questions themselves, the rate at which questions arise, the percentage of questions that clinicians attempt to answer, and the percentage of those questions that clinicians are ultimately able to answer (this section will focus on the first important outcome: identifying what, exactly, clinicians need to know during their daily work). Gorman's 1995 review[78] introduced a framework consisting of five major categories of information used by clinicians:

1. Patient data

⁵See Section 3.2 for more discussion of these sorts of studies

⁶ The specific questions examined in Gorman et al.'s 1994 study[76] were actually originally collected for use in a 1995 study by Gorman & Helfand[77]; the 1994 MEDLINE study represents a secondary use of the questions.

Type of information	Description	Examples	Useful sources
Patient data	Refers to a specific person	Medical history Physical exam Laboratory data	Patient, family, friends Medical record
Population statistics	Aggregate patient data	Recent patterns of illness Public health data	Recent memory Public health departments
Medical Knowledge	Generalizable to many persons	Original research Textbook descriptions Common knowledge	Journal literature Textbooks Consultants, colleagues
Logistic information	How to get the job done	Required form Preferred consultant Covered procedure	Local policy & procedure manual Managed care organization
Social influences	How <i>others</i> get the job done	Local practice patterns	Discussion with colleagues

Table 3.1: Gorman's framework of clinical information. Reproduced from [78].

2. Population statistics
3. Medical knowledge
4. Logistic information
5. Social influences

Table 3.1 describes each type of information type in more detail. While these categories represent a useful framework, they are quite general, and many studies have attempted to elucidate more specific data about clinical information needs. One influential example of this sort of study was published in 1999 by Ely et al.[79] Ely directly observed a random sample of 103 family doctors in Iowa as they conducted their daily clinical work. The investigators recorded all questions that arose during patient encounters. If a subject chose to pursue an answer to a question, the investigators recorded which resources the subject used and how much time they spent; if the subject did not follow up on a question, the investigators asked why. The investigators recorded 1,101 questions during the course of the study, and categorized them into a taxonomy of generic question forms. They found that the most common question topics centered around drug prescription, and that questions about other topics were rarely pursued. Table 3.2 lists the ten most common question forms observed during this study.

Generic question	# Asked (%)	# Pursued (%)	# Answered (%)
What is the cause of symptom X?	94 (9)	8 (9)	4 (50)
What is the dose of drug X?	88 (8)	75 (85)	73 (97)
How should I manage disease or finding X?	78 (7)	23 (29)	19 (83)
How should I treat finding or disease X?	75 (7)	25 (33)	18 (72)
What is the cause of physical finding X?	72 (7)	13 (18)	6 (46)
What is the cause of test finding X?	45 (4)	18 (40)	13 (72)
Could this patient have disease or condition X?	42 (4)	6 (14)	4 (67)
Is test X indicated in situation Y?	41 (4)	12 (29)	10 (83)
What is the drug of choice for condition X?	36 (3)	17 (47)	13 (76)
Is drug X indicated in situation Y?	36 (3)	9 (25)	7 (78)

Table 3.2: The ten most frequently-asked generic question forms identified by Ely et al. in [79]. Note that the percentage of questions that subjects actually pursued varied by question type: the subjects pursued answers to “action” questions (“What is the dose of drug X,” “What is the drug of choice for condition Y”) much more frequently than they pursued answers to “knowledge” questions (“What is the cause of symptom X,” “Could this patient have disease or condition X”).

A second study conducted by Ely et al. combined observational data from the 1999 study with data from a 1995 study by Gorman & Helfand [77]) to form a pool of observations covering 250 primary care physicians in Iowa and Oregon and amounting to more than 1,300 unique clinical questions. They then attempted to categorize the questions into a hierarchical taxonomy[80]. The taxonomy they used included five top-level categories of clinical question (diagnosis, treatment, management, epidemiology, and non-clinical), and proceeded through three increasingly narrow categories (see Table 3.3).

In 2007, González-González et al. published a study describing the results of their study of 112 Spanish family physicians and pediatricians. The investigators video-recorded four hours of each subject’s patient consultation. After each patient encounter, the subjects were asked to identify all clinical questions that arose during that patient’s session by stating them to the video camera. A panel of three clinician-researchers reviewed each subject’s video, and categorized the questions according to the taxonomy devised by Ely et al.[80]. They found that their subjects’ questions generally fit quite well into the taxonomy; however, they did extend the Ely taxonomy to include taxa for questions about administrative matters (“What

Code	Primary	Secondary	Tertiary	Quaternary	Generic Type
1.1.1.1	diagnosis	cause/interpretation of clinical finding	symptom		What is the cause of symptom x ? OR What is the differential diagnosis of symptom x ?
1.1.2.1	diagnosis	cause/interpretation of clinical finding	test finding		What is the cause of test finding x ? OR What is the differential diagnosis of test finding x ?
...					
1.2.1.1	diagnosis	criteria/manifestations			What are the manifestations (findings) of condition y ? OR What is condition y ? OR What does condition y look like?
...					
2.1.1.2	treatment	drug prescribing	how to prescribe	dosage	What is the dose of drug x (in situation y)? OR Should I change the dose of drug x (in situation y)?
2.1.1.3	treatment	drug prescribing	how to prescribe	timing	When (timing, not indication) or how should I start/stop drug x ? OR For how long should I give drug x ?
...					

Table 3.3: A subset of the taxonomy used in [80].

are the administrative rules/considerations in situation y ?)” as well as for issues of provider education (“I need to learn more about topic x .”). In contrast to Ely’s finding that the most common category of question related to drug prescription and dosing, González-González et al. found diagnostic questions (“What is the cause of symptom/physical finding x ?”) to be the most common category (35% of questions).

One important question is the rate at which questions arise during clinical care. Answers to this question vary widely from study to study. Coumou et al.’s review of the subject found that reported question-emergence rates ranged from 0.07–1.85 questions per consultation, and that the rates varied according to clinical setting (e.g. urban vs. rural, solo vs. group practice etc.) as well as by study design (survey vs. direct observation, etc.) and question definition[81]. The subjects in the González-González study seemed to come in towards the low end of that range, with 0.18 questions/recorded consultation[55], as do the subjects in Ely et al.’s 1999 study, who raised 0.32 questions per patient[79]. Osheroff et al.’s study of an inpatient teaching hospital recorded an average of five questions per patient encounter[82], but this included questions raised by instructors for purposes of evaluating trainees knowledge, which might explain its larger size relative to other studies.

The rise in popularity of evidence-based medicine (EBM)[83] has brought with it an increased focus on teaching clinicians to formulate clinical questions in clear and systematic ways. One very influential and popular model for this sort of clinical metacognition is the “PICO” framework. Users of this approach are encouraged to think of their clinical questions in terms of the *patient(s)* involved (“P”), the *intervention or exposure* under consideration (“I”) along with a *comparison* intervention or exposure (“C”), and the *outcome* of interest (“O”).[3] By expressing their clinical questions in this format, clinicians are forced to turn what can be very

abstract and unfocused “feelings” into concrete and (more importantly, in terms of EBM) *actionable* questions.

For example, consider an example given by Guyatt: a 35-year-old female patient, pregnant with twins, who is trying to decide between a planned vaginal delivery or a planned Cesarean section. In this case, the “patient” would be “term twin pregnancies,” the “intervention” would be a planned Cesarean section, the “comparison” would be a planned vaginal delivery, and the “outcomes” would be infant mortality and possibly maternal morbidity.

While this question formulation methodology may seem very straightforward, identifying and framing questions in this manner does not come naturally to many clinicians[5]. In a 2010 Cochrane review, Horsley et al. identified four interventional studies attempting to improve question formulation on the part of clinicians; while three of the four studies did show some improvement in question-formulation skills, the reviewers had concerns about bias and a lack of long-term effects on the part of the interventions [84].

3.2 When, how, and where do clinicians get answers?

As mentioned earlier, many studies of clinical information behavior begin by cataloging information needs, but also collect data on which clinical questions subjects chose to pursue, and how well they were able to find answers to those questions. The evidence is clear: clinicians do not seek answers to the vast majority of clinical questions. The clinicians described in González-González et al.’s 2007 paper pursued only 23% of their questions[55]. Other studies have observed somewhat higher numbers (Bennett et al.’s subjects reported investigating approximately 30% of their questions[85]; Ely et al.’s subjects pursued 46% of their questions), but the general pattern remains. This raises the question of why clinicians

pursue answers to some questions and not to others.

The defining characteristic of clinical information behavior is its compressed time frame. It is beyond cliché to state that clinicians are busy people, and, indeed, one of the major information-related difficulties reported by clinicians is the amount of time it takes to find quality answers[81]. When clinicians do end up seeking external answers, they do it quickly: Ely et al. found that, on average, clinicians spend less than two minutes seeking answers to any given clinical question[79], and González-González et al. found similar results in their 2007 study[55]. Busier physicians tend to ask fewer questions in the first place[86], and it seems safe to assume that they also pursue fewer questions.

The fact of these extreme time constraints has colored much of the work that has been done on physician information behavior. One highly influential model of physician information-seeking behavior is the “utility-choice” proposed by Connelly et al. in 1990[87].⁷ Essentially, Connelly’s model frames the question of whether and how to seek information in response to a clinical question as one of costs and benefits:

The decision [of] whether and where to seek knowledge is represented as a compromise between the conflicting goals of a need for information that reduces uncertainty, and a resistance to time, effort, or monetary expenditures. If the use of all knowledge resources is perceived to be more costly than is warranted by the value of the information expected from them, no additional knowledge will be sought. If more than one knowledge resource is perceived as having greater value than cost, the source that is perceived as having the greatest net value will be pursued.

⁷ Connelly was not the first researcher to look at clinical information behavior from this perspective (Huth’s paper in 1985[88] may have this distinction); however, Connelly’s nomenclature and model appear to be the most widely cited.

Logically, it follows that physicians will use the information source that they perceive to have the lowest costs (in terms of time, money, or frustration) and the highest benefits (in terms of reduction in uncertainty, authoritative answer, etc.). One influential study by Curley, et al.[1] surveyed 228 physicians about perceived costs and benefits of various information sources, and found that the significant determinants of source preference were the “cost” variables (availability, searchability, understandability, etc.) rather than the “benefit” variables (extensiveness, credibility, etc.). While this study is somewhat dated, its central findings may likely still apply today.

Of course, before clinicians make any decisions about *which* information source to use, they must first make the decision of *whether or not* to even pursue an answer to their question. In a 2005 study of 48 physicians, Ely et al.[89] found that subjects pursued answers to just over half (55%) of their questions; often, their reason for not pursuing a question was a lack of belief that they would find an answer. This reinforces Gorman et al.’s 1995 finding that that the most important factor affecting a physician’s decision of whether or not to seek an answer to a clinical question was whether or not the physician believed a definitive answer to exist[77].

Once the clinician has made the decision to seek information, they are immediately faced with a variety of options. A review conducted in 1995 by Verhoeven, et al.[2] found that physicians primarily consulted their colleagues, followed by books and journals, libraries, and finally “printed or online bibliographies.” Verhoeven also reported age-related differences in information behavior, with younger physicians being more likely to consult libraries than older physicians. A contemporaneous meta-analysis by Haug[90] found roughly similar patterns, although the Haug study put books and journals ahead of colleagues. Smith’s 1996 study[56] also cited the importance of colleague consultation, with the added dimension that many clinicians may in fact be seeking affirmation as well as information when

they go to a colleague with an information need.

Obviously, these studies pre-dated many of today's widely-used electronic medical reference products, such as UpToDate. A 2005 literature review by Coumou, et al.[81] looked at articles from 1992 through 2005, and found that the introduction of modern electronic resources did not change the underlying problem of clinicians having insufficient time and skill to answer many of their questions. On the other hand, a more recent study by Hider, et al.[91], conducted in 2009, found that many physicians and nurses made frequent use of electronic resources. However, they also reported approximately equal frequencies of search engine use and colleague consultation. Notably, the subjects in this study reported using Google more than any other online resource; the investigators also observed large differences between physicians and nurses in terms of which non-Google resources they used most often.

Once a clinician has decided to pursue a clinical question or information need, and has chosen an information source to use, the logical question becomes: how well are clinicians actually able to find answers to those questions they choose to pursue? There are really two aspects to this question. The first is what percentage of clinical questions even have answers waiting to be found, and the second is to what extent clinicians will be able to find the answers that *do* exist. Both aspects are important; a 1994 study by Gorman, et al.[76] found that many "native" (i.e., real-world) clinical questions encountered by practicing physicians could simply not be answered using the published medical literature, even when the research was done by trained librarians. Of necessity, however, most research has focused on the second aspect, and there exists a robust literature exploring how well clinicians are able use electronic resources to answer clinical questions, and results seem to vary considerably from study to study.

A 1998 review by Hersh and Hickam[6] found several studies showing that, in

cases where a clinician was unable to answer a clinical question using an electronic resource, the root cause lay in the clinician's search strategy rather than in the content of the resource. The subjects in a 2001 study by Alper, et al.[92] attempted to answer a set of 20 real-world clinical questions using 14 different clinical databases (STAT!Ref, MDConsult, etc.), and found that they were indeed able to find answers to most of their questions, but that it took them far longer than the previously-mentioned two-minute time span that practicing clinicians typically have in which to answer questions. The authors noted, however, that the speed issues were due at least in part to slow computer equipment and software rather than to the clinicians themselves, and it is reasonable to suppose that this state of affairs may have improved over the last ten years. Similarly, the medical students who participated in a 1996 study by Hersh[93] took an average of approximately 30 minutes to resolve a clinical question (either correctly or incorrectly). Again, though, the technology they were using was very primitive by today's standards. Furthermore, in 1996, querying electronic databases was something of an esoteric skill; today, it is a much more common activity, and it is not unreasonable to suppose that modern clinicians might be able to search more quickly.

This may, however, be an incorrect supposition. Several studies examining the actual search behavior of clinicians have noted that clinicians, like Internet searchers in general, tend to use relatively primitive search techniques. Herskovic, et al.[94] analyzed the raw query history for a typical day of PubMed search results, and while they did not limit themselves to queries run by clinicians or even queries on clinical topics, their findings are instructive. The queries included a median of three terms, and very few (11.2%) included operators or boolean terms. A more clinically-oriented study by Meats, et al.[95] surveyed the search logs of a meta-search-engine that retrieves results from numerous sources of medical information, including PubMed, the Cochrane library, etc. They found that most queries

included only a single term, and that approximately 12% included a boolean operator. Of this 12%, nearly all used the AND operator, with fewer than 1% using OR.⁸ One very early study by Haynes, et al.[96] found that while clinicians were (for the time period) favorably disposed towards searching MEDLINE, their performance lagged quite a bit behind that of trained medical librarians, to the extent that the clinicians retrieved half as many relevant results and 50% more irrelevant ones. That said, however, the subjects in the Haynes study did report that 47% of their searches affected clinical decisions.

In spite of the difficulties that clinicians have at using search systems, a number of studies have found that electronic resources are ultimately able to noticeably improve clinicians' performance. A 2005 study by Westbrook et al.[97] explored the question of whether online information retrieval systems actually *could* help clinicians to answer clinical questions. A panel of six clinicians created a set of eight clinical scenarios based on real-life cases; for example: "What is the best delivery device for effective administration of inhaled medication to a 5-year-old child during a moderately severe acute asthma attack?" Six of the scenarios had multiple-choice answers ("yes," "no," "conflicting evidence," and "I don't know"), and two of them had short one- or two-word answers. They enrolled 75 experienced Australian clinicians (including hospital-based doctors, family physicians, and clinical nurse consultants). Each subject reviewed the eight scenarios, and attempted to answer them unaided. Immediately after completing their answers, the clinicians were given access to a collection of six different electronic information sources (PubMed, an electronic version of the *Merck Manual*, etc.) and told to find and document evidence to answer each clinical scenario.

Westbrook et al. found that "pre-test" (i.e., without the aid of the information

⁸ It seems possible that many of the queries that Meats, et al. considered to be using the AND operator were, in fact, simply using the English word "and"—meaning that the users entering the queries did not realize that they were using a boolean operator at all.

systems), clinicians answered 29% of the scenarios correctly. Post-test (i.e., with the aid of the information systems), that number increased to 50%. While this is an encouraging result, the authors made sure to point out that 7% of their subjects incorrectly changed their answer after using the electronic resources— i.e., they had initially been correct in their response to the clinical scenario, but after using information retrieval tools, chose to change their response to an incorrect one. The authors consider this to be an example of “unintended consequences” resulting from the use of information technology in healthcare settings (as described by Ash et al.[98]), and point out the dangers of automation complacency (which Westbrook et al. define as the tendency of individuals to over-rely on computerized systems).

In addition to studying how well clinicians themselves are able to use electronic resources, numerous researchers have studied ways to increase clinicians’ utilization of librarians and other similar information professionals. In 2000, Davidoff and Florance[99] published a provocative editorial in *Annals of Internal Medicine* in which they called for the creation of a new medical discipline, that of the “informationist.” After enumerating numerous studies detailing the issues clinicians face in fully utilizing the medical literature, they announce:

We believe it’s time to face up to the fact that physicians can’t, and shouldn’t, try to do all or even most medical information retrieval themselves. In the current environment, that makes no more sense than it would for physicians to perform all or most of their own clinical chemistries, electrocardiography, computed tomography, and the like. Better they should focus their scarce discretionary professional time on reading, discussing, and reflecting in ways that truly deepen their conceptual and practical understanding of medicine than on the mechanics of finding, extracting, and synthesizing information from the published

literature.

Responding to this call, the NIH began hiring clinician-librarians to serve as informationists in its clinical research facilities. A 2010 work by Grefsheim, et al.[100] conducted a series of focus groups and observational studies on the NIH informationists, and found that they were having a positive effect on their clinical team members in terms of those members' ability to access electronic resources and follow up on clinical questions. Even better, the investigators also found increase in the clinical staff's *likelihood* of accessing electronic resources and following up on clinical questions.

One of the programs that inspired Davidoff and Florance's editorial is located at the Vanderbilt University Medical Center (VUMC), where, in conjunction with the university's Eskind Biomedical Library, features a "Clinical Informatics Consult Service," which "integrates librarians into clinical rounding teams"[101] and from which clinical staff may request informational support. This service is consistently rated very highly by VUMC staff.[101, 102]

In summary, there is a vast literature on the information behavior of clinicians; however, much of that literature pre-dates today's technological milieu. Today, most professional adults have at least some degree of computer-literacy, which in this day and age means having a greater degree of comfort and proficiency with search tools than would have been the case even ten years ago. However, although the technology may have changed, the fundamental challenges facing clinicians have not: put simply, their information needs are greater than ever, but the amount of time that they have available to meet those needs has remained constant (or has perhaps diminished).

We close this chapter by pointing out that virtually all of the research we have presented has been focused on the information needs and behavior of English-speaking clinicians in North America, the United Kingdom, Australia, and New

Zealand. There has been relatively little research done on the information behavior of non-native English-speaking clinicians, and much of what has been done has focused on Scandinavia and Western Europe. One notable exception is the work of Martinez-Silveira and Oddone[103], which looked at the information behavior of medical residents in Bahia, Brazil. The medical residents' specific information needs were not notably different from those of North American clinicians. Drug therapy information and diagnostic questions were the two most frequently-experienced information needs. A surprisingly high proportion of the subjects (41%) reported using a medical library "frequently," although most of the subjects (81%) reported that they performed their own database searches (rather than relying on a librarian). MEDLINE was the preferred information source; only half of the self-searchers reported using the LILACS database of regional medical literature. Interestingly, virtually none of the self-searchers reported having had any training whatsoever in how to search the medical literature; only one subject reported having had training from a librarian. Just as their North American colleagues did, most of the subjects reported using relatively simplistic search strategies.

It should be noted that Martinez-Silveira and Oddone made no mention of whether or not their subjects experienced any language-related limitations. However, a 2007 study by Gonzalez-Gonzalez, et al.[55] that looked at Spanish clinicians did note certain linguistic limitations among their subjects. This study followed a similar pattern to several of the other studies we have discussed in this chapter; the investigators recorded a large sample of clinicians ($n = 112$) engaging in patient interactions, recorded any clinical questions that arose, and coded the questions according to Ely's taxonomy[80]. The investigators also looked into which questions the clinicians chose to follow up with, and how they did so. They found that fewer clinicians used the Internet to address the questions than they had

initially expected, and attributed this to the fact that 72% of the clinicians did not have Internet access in their offices. The investigators were, however, surprised to find that very few of the subjects that *did* use the Internet to research a clinical question were using secondary electronic resources (such as Epocrates, STAT!Ref, etc.). The investigators ultimately attributed this finding to the fact that these resources are published in English.

Considering the large body of evidence suggesting that clinicians will follow the “path of least resistance” when it comes to information sources, and considering further the extreme time limits under which clinicians work, is it any wonder that the clinicians studied by Gonzalez-Gonzalez would eschew resources that forced them to interpret a foreign language, even if it were one with which they were reasonably comfortable? Clinicians here in the US, working in their native language, have no shortage of barriers preventing them from making full use of the numerous information resources at their disposal; imagine if, in addition to overcoming these existing barriers, they also had to grapple with a language barrier? And yet this is the position in which most of the world’s clinicians find themselves. This dissertation describes a system designed to help address this issue, as well as an evaluation methodology aimed at assessing the effects of such tools on clinical users.

Chapter 4

Cross-Language Information

Retrieval and Multilingual User

Interfaces

Chapter 2 discussed some of the challenges facing non-native English-speaking biomedical researchers and clinicians who need to make use of a body of literature that is predominantly in the English language. These issues are, of course, not exclusive to these groups of people; anybody who needs to access information in a foreign language will run into many of the same difficulties. This problem is as old as human language itself, though the advent of the Internet— and the resulting widespread increase in communication across geographic and language boundaries— has in recent years made the problem more acute.

Not surprisingly, computers and other communication technologies have long been looked to as potential solutions to this problem. Automated translation of text from one language to another was one of the first non-numerical applications proposed for digital computers[104],¹ although in those days it was referred

¹ Note that even before the invention of the digital computer, various inventors were working on mechanical means of translating text from one language to another.[104]

to as “mechanical translation” rather than the more modern “machine translation.” Early efforts in the field were primarily focused on translating between Russian and English, with a goal of enabling American scientists and military planners to keep abreast of Soviet publications. The field experienced its share of advances and retreats through the mid-twentieth century (many of which are documented by the collected papers in Nirenburg, et al.’s “Readings in Machine Translation”[105]), and throughout the 1970s suffered serious funding difficulties following an influential US government report that not only highlighted the technical limitations of what were then the state-of-the-art systems, but also questioned the fundamental validity of the field as a whole[106, 107]. While machine translation systems did increase in quality over time, their performance was limited by the fact that they were primarily “knowledge-based” or “rule-driven” in nature. Such systems generally translated text from one language to another using sets of explicitly-encoded linguistic rules, and relying on human-curated databases of factual knowledge and dictionaries of terms. Two well-known examples of such systems are the venerable SYSTRAN[108] and PAHO[109] systems, both of which are still widely used (SYSTRAN for general-purpose commercial translation, and PAHO specifically for translating journal articles from English to Spanish and vice-versa).

In the 1990s, however, increased computer processing power and data storage, combined with new mathematical techniques, helped lead to the development of *statistical* machine translation (SMT). The first well-known SMT system was described by Peter Brown and his team from IBM[110] in 1990; more followed. Rather than relying on explicit rules, dictionaries, or language parsing techniques, SMT systems are “trained” on large corpora of text in both the source and target language, and automatically *learn* how to translate words and phrases. Classical SMT relies on *translation models* that map text from one language to another, and *lan-*

guage models that indicate how statistically likely or unlikely a possible mapping may be given its context. Early systems required that their corpora be very closely aligned, often at the sentence level, i.e., that each sentence in their source language's corpus have a corresponding equivalent sentence in the target language's corpus. This is known as a "parallel corpus," and as one might imagine, such corpora are not always easy to find. Over time, MT researchers have developed systems that can be trained on "comparable corpora," whose contents do not need to be made up of 1:1 sentence mappings.[111]

While machine translation can help users to read content written in a foreign language, on its own it does little to help them find such content. Journalists, patent examiners, individuals wishing to obtain technical documentation, private citizens seeking to read local accounts of foreign events—nearly everybody has, at one time or another, found the need to locate information in another language. The field of cross-language information retrieval (CLIR) is a sub-field of information retrieval focused on helping "searchers find documents that are written in languages that are different from the language in which their query is expressed"[112];² Oard's 1997 survey of the field is comprehensive and insightful, if somewhat dated at this point.[114] Machine translation and CLIR are closely related; many CLIR systems make direct use of machine translation technologies, and even those that do not use MT directly often rely on some of the same theoretical underpinnings as MT.

Furthermore, a common use model for CLIR is one wherein subjects will ultimately rely on a translator of some sort (either human or computer) in order to make full use of the information they retrieve using a CLIR system[113]. This is certainly a useful model, particularly in the context of CLIR involving Asian languages[115, 116, 117]; however, it must be kept in mind that there are many use cases for CLIR by "polyglot" users who may, to various extents, read the docu-

² Douglas Oard notes that CLIR has sometimes been referred to, perhaps uncharitably, as "the problem of finding people documents that they cannot read." [113]

ments that they are ultimately retrieving[118, 119]. These users represent a large potential user pool for CLIR systems, and will undoubtedly have different needs regarding user interface design than will monolingual users.

There are three basic approaches to CLIR. A system may attempt to translate the user's query, and then use standard information retrieval techniques to search across a document collection for articles matching that query; or, the system may translate the document collection into the user's language, and run the original query as-is. The third approach is actually a family of approaches, members of which typically take the form of some sort of hybrid of the first two approaches, often involving intermediate representations for both query and document collection. As one might imagine, this is not an easy task. Most machine translation systems are designed to translate "natural" text, and the queries that users enter into search engines are often anything but natural. Furthermore, until relatively recently, machine translation was both computationally and financially expensive, and as such was not a practical option for many ad hoc search scenarios.

As a result of this, CLIR researchers have devoted considerable attention to different methods of query translation (see Grossman & Frieder's chapter on CLIR for an excellent overview of some of the technical aspects inherent to this process[120]). These techniques can work quite well; as far back as 1998, Ballesteros and Croft's CLIR system was able to achieve more than 90% of the performance of an equivalent monolingual system[121], and Oard and Diekema reported similar results around the same time[122]. Of course, one could ask what this level of performance means in the context of a user needing to actually *use* a CLIR system to obtain foreign-language information. To address this question, we must turn our attention to the matter of *evaluation*.

For a variety of reasons, information retrieval evaluation in general tends to be heavily "system-oriented," often hewing closely to the evaluation paradigm be-

gun with the pioneering Cranfield experiments[123] and perpetuated by the TREC evaluation campaigns[124, 125]. Oard, et al. describe this model in an admirably succinct manner:

The typical way of evaluating ranked retrieval effectiveness is to obtain a collection of documents that are representative of those that would be searched in the actual application, create a set of queries that are representative of the way searchers are expected to express their interests in specific topics, somehow establish the relevance of each document to the topic represented by each query, and then compute a measure that reflects the density of relevant documents near the top of the list.[126]

In a standard IR evaluation, then, one's system attempts to satisfy each such query from the documents included in the collection, and its performance is operationalized as a function of whatever metric the investigator chose. This model, while useful, leaves the end user almost completely out of the picture, and can lead to a somewhat myopic focus on improving certain easily-measured aspects of system performance while simultaneously ignoring other, less-easily-measured, aspects of the system— for example, the extent to which end users are actually able to *use* the system itself. Numerous researchers have pointed out that information retrieval has always had, and will always have, a large human component[127]; final relevance judgments are ultimately made by humans, after all, and the fundamental reason that IR systems exist is to satisfy human information needs. Additionally, several authors have found that system-oriented metrics such as precision and recall do not necessarily correspond to user satisfaction or performance[128, 129, 130, 131].³

³ None of this is to say that system-oriented evaluation is without value; however, it is important to keep in mind that it is just one necessary step out of many rather than an end in and of itself. The

With the increased use of electronic information retrieval systems by “end users” (described by Marchionini in what he refers to as a “somewhat distorted but often-used sense” as being “everyone except professional intermediaries”[133]) in the late 1980s and early 1990s, the sub-field of *interactive information retrieval*— that is, information retrieval with an explicit human element— began to gain prominence, as IR researchers recognized the increased importance of the human element of their work, and found that evaluations that ignored the interactive aspects of retrieval was becoming increasingly less useful.[124, 134, 135, 136]⁴ The addition of an “interactive” track to TREC in 1994[125, 126] provided a formal venue for such research, and aided in the development of standardized test collections and experimental protocols.

From an evaluation standpoint, however, the problem is that designing a “black-box” experiment to measure an information retrieval system’s precision or recall is simple; as mentioned, the field has a rich history of experimental protocols, test collections, conceptual models, etc. from which to draw. Evaluation of interactive information retrieval systems is a much younger discipline, and faces numerous experimental challenges. Robertson summarized the central difficulty of the field thusly:

The conflict between laboratory and operational experiments is essentially a conflict between, on the one hand, control over experimental variables, observability, and repeatability, and on the other hand, realism.[124]

The question, of course, is how to maximize realism in our research while minimizing the evaluation framework outlined by Stead, et al.[132] describes how different levels of evaluation feed into and build off of one another.

⁴ Note that interactive studies were not *completely* unheard-of before this time; both Kelly[137] and Petrelli[118] refer to Salton’s work in the early 1970s[138] as the earliest family of IR studies that explicitly examined interactive effects, and Kelly also points out that even some of the earliest Cranfield reports made mention of presentation issues as well as recall and precision.

imizing the damage done to observability, repeatability, control over variables, etc. The most common pattern for interactive information retrieval studies appears to be Borlund and Ingwersen call a “simulated work task situation,”[139, 140] in which actual end users are enrolled as subjects, and are made to use the system (or systems) under investigation perform a some simulated task— typically, attempting to address an information need, which may be either synthetic (generated by the investigators) or may come from the subjects themselves. The outcomes from this sort of study are typically the subjects’ queries themselves, the documents that they found relevant to those queries, or some other metric of task completion or understanding. Kelly, in her exhaustive overview of IIR evaluation, refers to this style of experiment as an “archetypical IIR study”[137], and it is a common pattern for evaluation task, including those found in the “Interactive” track of TREC. Borlund has written several excellent overviews of studies conducted using just such an experimental model.[140, 141]

IIR researchers may choose from many different user-oriented metrics. In “Information Seeking in Electronic Environments”[68], Gary Marchionini describes several different methodologies for evaluating search performance. One 1988 study, by Wang, Liebscher, and Marchionini attempted to compare the performance of users under two different hypertext search strategies.[142] The specific research question they were attempting to illuminate is not relevant here; what *is* important is that they used a variety of functional measures as dependent variables: “success,” number of moves made, total time to complete searches, and the number of articles and screens viewed. They also examined general satisfaction among their users, as measured by “ease of use,” “speed of use,” and “frustration level.” Another formative study mentioned by Marchionini measured the number of articles that users chose to examine, the number of queries users executed, and the time that users took to execute searches[143].

These metrics remain popular, and are certainly useful. However, since using an information retrieval system is rarely an end in and of itself, these measures may perhaps be best thought of as “process measures”⁵ that hopefully represent meaningful proxies for some other, more fundamental outcome of interest. Indeed, some investigators have gone beyond simply looking at “process” measures of IIR (documents selected, queries entered, etc.) and investigated the feasibility of using “outcome” measures— i.e., they have attempted to directly measure how well a system solved a user’s actual information need.

One oft-cited example of this is a study by Hersh, et al., in which the authors investigated metrics of search performance based directly on information acquisition[145]. Their study presented medical students with a test consisting of a variety of clinical questions (e.g. “How is the organism which causes Rocky Mountain Spotted Fever transmitted?”). After the students completed the test, they chose the five questions whose answers they were least confident about. They then proceeded to use one of two IR systems to attempt to find definitive answers for the five chosen questions. The investigators found there to be a measurable increase in information gained (as measured by the number of correct answers pre- and post-IR) and concluded that, at least in analogous situations, metrics based on this technique could be viable for assessing the performance of IR systems.

However, this study was limited in scope and the observed effect was sensitive to the specific questions used to measure information gain. For example, the authors mention that one particular question’s wording seemed to cause confusion to the study participants and may have affected the magnitude of the observed effect.⁶ The general question of evaluation design and metric selection for interactive information retrieval remains an active area of research.[146, 147, 148]

⁵ We refer here to the concepts and terms from Donabedian’s “structure-process-outcome” model[144].

⁶ These sorts of issues are quite common in IIR studies, and highlight the difficulties inherent to any research involving human subjects.

4.1 CLIR Evaluation

The sub-field of cross-language information retrieval has followed the pattern of its parent field in that the field's evaluative focus has generally focused on studying system-oriented attributes rather than user-oriented or interactive elements. Gey, et al. describe several different CLIR evaluation campaigns, illustrating the high importance that the field's practitioners place on evaluation; however, almost none of these campaigns included user-centered evaluation.[149] Furthermore, Grossman & Frieder's chapter on CLIR devotes several pages to evaluation, none of which mention end users. During the late 1990s, however, several researchers did begin to formally evaluate usability issues in CLIR systems.

One of the first quantitative evaluations appears to have been that of Ogden & Davis, which found that English speakers with no knowledge of German were able to use a CLIR system to perform relevance judgments of German-language documents, and were able to do so with an accuracy competitive with that achieved by German-speaking TREC assessors[150, 151]. Their CLIR system used a simplistic dictionary-based approach to query translation, but notably used the venerable SYSTRAN machine translation engine to translate search results, which apparently was sufficient for the subjects.

These studies were, in many respects, very "traditional" in that they depended heavily on TREC assessments and focused on relevance judgment quality as the primary outcome. Other interactive CLIR⁷ studies have been more "task-oriented" in nature. Since novel CLIR systems typically feature novel user interface elements, interactive CLIR makes for a particularly rich context for research into human-computer interaction in general. As such, some of the studies we will discuss below focused more on how well (or poorly) subjects responded to various

⁷ Defined by He, et al[152] as an "iterative process in which searcher and system collaborate to find documents that satisfy an information need, regardless of whether they are written in the same language as the query."

aspects of the user interface than on outright retrieval performance.

One very representative study is that of He, et al.[152]. They describe their participation in the 2002 CLEF⁸ interactive task, in which subjects had to identify articles that were relevant to four different topics from a collection of German-language documents, using English-language queries. The subjects in He's study used two different search interfaces: one interface translated their queries in a fully automatic fashion, and another that provided significant manual control over the query translation process (referred to by He as "user-assisted translation"). The order in which subjects addressed each topic, and which interface they used to do so, was varied systematically to attempt to balance out any topic/system interactions.

The user-assisted translation interface provided several different kinds of translation support to the English-speaking user, primarily focused around enabling the user to make informed choices about how each word or phrase was to be translated. For each possible translation, the interface provided "back translations" (i.e., English translations of the initial German translation, which can be useful for diagnosing word-sense disambiguation issues) as well as "sentence-in-context" views, in which the user was shown a sample sentence from a search result containing one of the possible translated query terms.

As with many studies involving human subjects and information retrieval, their results were mixed.⁹ In general, however, they found that their subjects were able to make more accurate document selections when using the interface with user-assisted query translation, although the magnitude of this effect varied greatly by topic. Subjects performed more query iterations for the topics in which they were using the user-assisted query translation interface, and many of the sub-

⁸ One of the major venues for CLIR evaluation has been the annual Cross-Language Evaluation Forum, which started life as a track within TREC and eventually split off into its own series of evaluation campaigns. See Peters, et al.[153] for an typical collection of CLEF-related publications, and the CLEF website (<http://www.clef-campaign.org/>) for further CLEF-related information and a complete CLEF bibliography.

⁹ See [137] and [134] for numerous examples of other such studies.

jects seemed to appreciate having more control over the query translation process.

Another CLEF interactive task, this one from the 2008 CLEF, gave rise to several informative experiments[154]. That year's task centered around FlickLing[155], a multilingual interface designed to enable users to retrieve images from the Flickr photo-sharing service¹⁰ annotated in languages different from the one used in their query. So, for example, an English-speaking user could use FlickLing to search for images annotated with "car" and retrieve images annotated with "coche," Spanish for "car." The interface also supports several different secondary search tasks designed to a) encourage additional interaction with the system by the user, and b) solicit additional user-provided translations and annotations for images. FlickLing's secondary purpose, after serving as a search tool, is the generation of log files of user interactions with the system, which may then be mined to learn about how end-users interact with multilingual search interfaces. For the 2008 CLEF interactive task, FlickLing's operators obtained search logs representing approximately 5,000 searches from a linguistically diverse population of more than 200 users (for whom the task providers had a certain amount of demographic data).

Using these search logs, participants in the 2008 CLEF interactive task were able to study various aspects of how users chose to use the FlickLing system. For example, Vundavalli found that users, given the choice, seemed to prefer to search in their mother language, and that users reported feeling more confident when searching in their mother language[156]. When users chose to use interfaces in non-mother languages, they preferred to search in languages in which they had "active" skills (typically considered to be speaking and writing) as opposed to unknown languages or languages in which their skills were primarily "passive" in nature (listening or reading). Users were more likely to reformulate their queries multiple times when working in their mother languages than they were when

¹⁰<http://www.flickr.com>

working in second languages. Vundavalli noted that the more successful users tended to reformulate their queries frequently (as opposed to exhaustively reviewing multiple pages of results), suggesting that systems that facilitate query modification could be highly useful in multilingual contexts.

Another relevant study from the 2008 CLEF interactive track was that of Tanase & Kapetanios[157], who focused on user-contributed translations. The FlickLing system offered users the chance to contribute their own translations for photo annotations in the event that they were unsatisfied with the system's translations, and enabled users to build a "personal dictionary" of annotation translations. These interactions were included in the query logs used for CLEF 2008. The authors investigated the extent to which a user's degree of confidence with a language affected their usage and creation of personal dictionary entries. They observed that users were generally reluctant to use the system's assisted query translation feature, and that this reticence was greater when the user was unfamiliar with the target language. Instead, users typically preferred fully-automatic query translation.

While the CLEF interactive track has, over the last ten years, served as a valuable source of interactive CLIR research, it has certainly not been the *only* such source. One important body of research is that produced by Petrelli and colleagues at the University of Sheffield, in the UK, and at the Swedish Institute for Computer Science, in Stockholm, Sweden. Their "Clarity" CLIR system has been influential not simply for its technical attributes but also for the manner in which it was developed. The team undertook a well-documented user-centered design process[158, 159, 160] that involved questionnaires, videotaped direct observations, and interviews, all of which were designed to elicit user needs[159]. These activities revealed that some of the developers' central assumptions about how their users addressed CLIR tasks had been incorrect. The developers had initially

thought that their users would desire fine-grained control over the query translation process;¹¹ this was based on what they felt to be “best CLIR practice.”[161] To the investigators’ surprise, however, they found that most of their users were unconcerned with the details of the translation so long as the final search results were adequate.

After the initial user studies, the group designed prototypes of two different interfaces to their system. The first interface provided direct user control over the query translation process (the “supervised” interface), while the second interface handled query translation completely automatically (the “delegated” interface). Following the framework described by [141], the investigators used a “simulated work task” design with actual end users to investigate which interface was superior. Their report[118] is enlightening. In terms of recall and precision, subjects objectively performed somewhat (but not significantly) better under the “supervised” interface; however, the subjects themselves tended to prefer the “delegated” interface, saying that it required “less effort” and was “easier... and quicker”.

Some subjects did prefer the supervised mode, appreciating the extra control. The authors note, however, that other subjects seemed to rate the supervised mode more negatively as a result of being able to actually see the system’s translations, which were not shown under the delegated mode. Subjects expressed significant frustration at the system’s limited translation capabilities— but only when they could *see* the translation itself! The authors noted, though, that in spite of the questionable quality of the system’s translations, some users found them to be useful aids to query re-formulation. The final interface as implemented by Petrelli combined the two interfaces such that the query was automatically translated, but, at the top of the results screen, users were able to view and modify the translated query. The authors felt that this addressed the issues of speed and effort while

¹¹ Hardly surprising, given the CLIR field’s heavy focus on minutiae of query translation.

preserving the benefits of being able to manually review translated queries.

Another valuable series of experiments is that described by Oard, et al.[126] in a multi-site study involving researchers in Spain, the UK, and the United State¹². The authors performed several experiments, one of which compared user performance when using “gloss” translations (wherein a bilingual term list is used to translate words or phrases from each result) against full machine translation. Gloss translation is easier to accomplish from a technical standpoint, particularly in the case of language pairs where full machine translation systems do not exist (e.g., Basque, most African languages, many Indian languages, etc.), and previous work by Oard and Resnik had found that subjects using a glossing system were able to perform cross-language document selection reasonably well[115]. In addition to the glossing-vs.-MT experiment previously described, the authors also performed experiments comparing SYSTRAN machine translation of results against a custom phrase-based translation approach, and another experiment involving native English speakers making relevance judgments of English-language articles as well as French-language articles that had been machine-translated into English.

As with many interactive CLIR studies, the experiments described in this article suffered from low sample size (each site’s experiments used eight subjects) and each group observed a great deal of between-subject and between-topic variation. Overall, however, their results were informative. They found that subjects performed better when results were fully machine-translated than when they were simply glossed, although the group experimenting with more sophisticated phrase-based translation techniques found that there was some reason to believe that phrase-based approaches could augment machine translation, and could work tolerably well in the absence of machine translation.

¹² It should be noted here that CLIR research in general, and interactive CLIR research in particular, is much more widespread in Europe and Asia than in the United States. The reason for this is probably related to the higher levels of linguistic diversity found in those parts of the world as compared to that seen in the United States.

Lopez-Ostenero, et al.[116], however, found the opposite result. In their 2005 study, they evaluated a system that used noun-phrases as the unit of phrase-based translation for both query and document translation, and compared it against a full-MT solution. Their phrase-based approach attempted to identify key phrases, and used a dictionary-based approach to translate those phrases into “phrase-based pseudo-summaries”. They found that, overall, their subjects seemed to prefer this approach to one that simply used MT to translate the document itself, and that there appeared to be no decrease in accuracy— in fact, users seemed to have significantly higher recall when using the phrase-based summary system than when using a SYSTRAN-based system. The authors also investigated the use of their phrase-based translation technique as tool for assisting users in query construction. They found that this system decreased initial query formulation time by 85%.¹³

In addition to studying how well CLIR systems enable users to perform document selection, several researchers have studied CLIR systems’ effects on how *quickly* subjects can perform such tasks. Suzuki, et al. conducted an evaluation in which 64 Japanese-speaking subjects performed a document selection task under three different interfaces.[117] The task consisted of identifying a relevant article from a list of ten possibly-relevant articles, for eight topics. Subjects were able to select up to three articles per topic, and were assigned a “point” if any of their selections included that list’s relevant article; since there were a total of eight topics, each subject could have anywhere from zero to eight points. The primary outcomes were the number of points earned as well as the amount of time each subject took to perform their selections for each topic.

¹³ The fact that the two studies found opposite results is intriguing, but not entirely surprising. The two groups were studying different systems, with different subjects, and took place nearly ten years apart. One possible explanation could be technological in nature: it is possible that Lopez-Ostenero, et al. used a newer and better-performing version of Systran than did Oard, et al., and as such had higher-quality phrase translations.

Under the first user interface condition, subjects viewed the result list in its original English; the other two conditions involved two different variations on dictionary-based translation from English to Japanese. The authors found small (but significant) differences in their subjects' accuracy scores between the various interfaces, but reported a dramatic difference between interfaces in the amount of time it took the subjects to perform the selection task. When using one of the two Japanese-language modes, subjects took approximately 30% less time than when they were using the English-language interface (an average of 145 seconds in English vs. approximately 105 seconds under the Japanese modes). Interestingly, Suzuki, et al. did not find English proficiency to be significantly related to any of their outcomes, although they did observe that subjects with more self-reported experience with information retrieval systems tended to make slightly more accurate selections than less experienced subjects. As with language proficiency, however, they did not find any statistically significant effect from IR experience.

Hansen & Karlgren have reported similar results from their experiments on Swedish-speaking subjects.[162] Out of a somewhat complex series of evaluations involving a variety of different use scenarios, they found that their subjects— most of whom reported relatively high levels of English proficiency— consistently were able to perform a relevance judgment task more accurately and more quickly when reviewing Swedish-language newswire articles¹⁴ than when reviewing English-language articles on the same topic. Subjects took an average of 27 seconds per English-language document and an average of 20 seconds per Swedish-language document. The assessment time did vary significantly from topic to topic, however, as is commonly the case in IIR studies.

¹⁴ It should be noted that, for this study, the authors were able to make use of existing topics and document collections, in this case those from the 2002 CLEF campaign. This highlights the value that researchers derive by participating in evaluation fora such as CLEF— not only were Hansen & Karlgren able to leverage existing resources (thereby saving time, money, and aggravation), it might in principle be possible to compare their data and results with those from other studies using the same CLEF resources.

These studies have all been largely focused on interfaces for information retrieval systems of one sort or another. However, IR is not the only use case for multilingual interfaces. For example, Ogden, et al. described a multilingual instant-message interface that used machine translation to enable users who spoke different languages to transparently exchange messages with one another.[163] The authors had previously shown[164] that existing machine translation software could be used successfully for this purpose, but had noted that it was far from perfect and that these imperfections led to slower and rougher communication, as users had to devote time and resources to repairing flawed translations.

This study paired native speakers of Japanese and English via a multilingual instant messaging (IM) interface. Each pair then undertook to perform a collaborative photo identification task in which one subject viewed an image while the other subject asked questions about the image's content, which the image-viewing subject then had to describe. All communication took place via IM communication. Some subjects used an IM interface that showed only one MT system's translation of each participant's messages, whereas other subjects used an interface that displayed translations from several different MT systems. The study also featured a control group whose subjects were pairs of native English-speakers communicating in English. The experimental outcomes were the percentage of correct image identifications, time taken per image, and the number of messages exchanged per image.

Neither interface mode turned out to be significantly "better" in terms of any of those metrics. Although the subjects using interfaces featuring multiple translations did take a bit less time and exchanged slightly fewer messages (with no meaningful change in accuracy), the authors were hesitant to attribute those differences to the effect of having access to multiple translations. As part of the study, independent judges rated the performance of the different MT engines used, and

found differences in translation quality between engines. Furthermore, when looking at the subjects' performance solely in light of which translation engine they were using, the authors found differences between engines that were similarly in scale to the differences they had noted between interface modes. As such, they feel that there is a possibility that the true cause of the between-interface-mode effect may have been differential translation quality, and conclude that "We still don't know if providing multiple translations has a benefit over just ensuring that at least one good translation service is provided."

Moving even further afield from traditional IR research, there exists a very large body of literature surrounding the general problem of "foreign language reading" in electronic environments. Much of the work in this field is conducted by researchers in foreign language (often referred to as "L2," even in instances where the language in question is a student's third, fourth, fifth, etc.) acquisition rather than information retrieval, and there is, sadly, relatively little overlap between the fields. Many of these studies have investigated various features that can be added to interfaces designed for use by learners of foreign languages, such as integrated bilingual dictionaries, textual annotations, etc. Since many of these interface elements may have application to CLIR systems in general, the findings of these studies may be relevant to the field of interactive CLIR.

One early and representative study is that of Aust et al. from 1993.[165] The authors conducted an early evaluation of the effects of hypertext annotations in L2 reading, in which intermediate-level English-speaking undergraduate students of Spanish were presented with a short Spanish-language passage as well as a reference tool consisting of either a paper or electronic dictionary. The authors randomly assigned their subjects to one of two reading environments (electronic or paper) and one of two reference tools (bilingual or monolingual dictionaries), for a total of four groups. In the electronic reading environment, subjects were

able to “click” on individual words to bring up either monolingual or bilingual definitions; under the paper-based environment subjects were left to contend with traditional monolingual or bilingual dictionaries. The authors measured the total time taken to complete the reading as well as the number of definitions looked up by each subject. The authors also administered a reading comprehension test to each subject based on preposition recall.

In what seems to be a recurring pattern in the literature, the authors found that the subjects who were in the “electronic” group made heavier use of the reference tools than did the subjects in the “paper” group. They also found that subjects with access to bilingual dictionaries looked up more words than subjects with monolingual dictionaries, and did so in a more time-efficient manner. Furthermore, the authors found a significant interaction effect between dictionary modality and reading environment on frequency of dictionary usage that indicated that the combination of bilingual reference and hypertext modality was more successful than other combinations of the two.

However, in keeping with the aforementioned pattern, the authors found no significant difference in reading comprehension between the two groups. The “electronic” group had lower mean comprehension scores under both bilingual and monolingual dictionary conditions, but the standard deviations were quite large and prevented any inference. This may have been due to the study’s small sample size: there were forty subjects in total, meaning that each condition only had ten subjects. However, in the years since this study, there have been numerous other studies (some of which we shall be discussing shortly) showing that hypermedia annotations have either negative effects or no effect on reading comprehension, and very few that have shown a positive effect, so it is quite possible that a larger sample size would not have changed the authors’ conclusions.

It is important to note that, in follow-up interviews with their subjects, the

authors noted that subjects' impressions of the hypermedia bilingual dictionary were all favorable, even though it did not seem to have any serious effect on their reading comprehension. So, we find ourselves in an interesting position: users clearly use hypertext glossaries when they are available, and have positive things to say about them... but hypertext glossaries do not seem to have a significant effect on primary outcome measures such as reading comprehension. However, the authors point out that other objective measures— including some that we might think of as being “process” measures, such as total time to complete a reading task— *were* affected in positive directions by the bilingual hyper-references. What does this tell us? It tells us that, in evaluating the design and impact of linguistic support features of user interfaces, we would do well to adopt more “holistic” views of the search process that take multiple measures and their impacts on the user's experience into account, an opinion shared by a number of researchers in the field[158, 118, 122, 137].

Many studies have investigated linguistic support features that are more ambitious than simple hypertext definitions. In a 1997 study, Davis describes one such experiment,¹⁵ involving an interactive application designed to assist students learning the French language in reading a textual passage from Ferdinand Oyono's “Une vie de boy.”[166] The authors constructed a reading application rich in domain-specific language support: bilingual glosses, contextual information derived from the text (such as interactive family trees for some of the story's characters, etc.), cultural notes, reference photos, etc. The authors' system also logged user interactions with the system in an attempt to track readers' metacognitive processes, and evaluated their subjects' reading comprehension with a follow-up test.

To the author's surprise, they found that while many subjects reported find-

¹⁵ In addition to being an interesting experiment, this study provides an excellent discussion of pre-1997 literature on the topic of supported L2 reading.

ing the various interactive features *helpful*, the subjects made very little use of any of them except for the bilingual glossing, which the subjects *did* use heavily. Furthermore, the authors found very little evidence that use of the various interactive features aided subject reading comprehension¹⁶, although they did point out that, had the subjects made more use of the various features, their comprehension might have been higher. This finding— that multilingual support features have little to no discernible effect on user reading comprehension— is one that repeats itself over time, although it is difficult to say whether this says more about the utility of such features or the difficulty inherent to measuring reading comprehension in a reliable and valid way.

In addition to describing an interesting experimental finding, a 2005 article by Sakar and Ercetin[167] also provides a thorough overview of several notable studies on the effectiveness of hypermedia annotations in an L2 reading task. The authors note that past research has had mixed findings with regard to what sorts of annotations are most preferred by users and what the effects of those annotations are on reading comprehension, and proceed to describe an experiment involving a hypermedia reading environment consisting of an English-language text and a variety of multimedia (audio, video, photographic, textual, etc.) annotations. Users could click on individual words or passages to bring up a relevant annotation, and each page contained several contextual annotations as well. The authors' study used Turkish students of "English for Academic Purposes" and directly tested the subjects' use of annotations as well their reading comprehension. In order to control for language ability, the authors only included subjects who scored within a relatively narrow range on the Oxford Placement Test. Although there were some problems with the authors' study design, their results (particularly their qualita-

¹⁶ The author's description of their study's conclusion is admirably pithy: "Perhaps the most striking finding in this study was the contrast between the overwhelmingly positive feelings toward the computerized gloss and the lack of any clear evidence that the program had, in fact, enhanced comprehension."

tive ones) remain informative.

They found a significant negative correlation between reading comprehension and annotation use— i.e., the more a subject made use of the system’s annotations, the lower their ultimate level of reading comprehension. Interestingly, upon stratifying their data by annotation type (audio, video, etc.), the authors found that certain annotation modalities were associated with decreased reading comprehension while others were not. For example, use of annotations that discussed word pronunciation or those that presented videos were strongly correlated with decreased reading comprehension, whereas purely textual annotations were not.

The qualitative results of Sakar and Ercetin’s study were somewhat at odds with their quantitative observations. As in other studies of hypermedia annotations, this study’s participants generally regarded the annotations as useful, although they found some modalities (video, etc.) much more so than others (audio annotations demonstrating proper word pronunciation). The authors speculated that this may be a case of users expressing a positive opinion about a particular feature more because they find it interesting than because they found it to be actually useful. The authors at one point speculate that the negative correlation they observed between annotation use and reading comprehension represented a causal relationship, and propose an explanation derived from a cognitive theory of multimedia learning: that the extra information from the annotations “overloaded” the subjects and thereby impeded their reading comprehension.

While this explanation is certainly a reasonable one, the authors of this study pointed out several important limitations that prevent us from drawing any firm conclusions about the utility and effects of hypermedia annotations. The most important of these limitations was the lack of a control group of students. Furthermore, as is the case with many studies of user interface features, it is difficult to disentangle which parts of the results relate to specifics of the authors’ implemen-

tation of hypermedia annotation from those parts that relate to the general *concept* of hypermedia annotation.

The last such article we wish to discuss here is that of Akyel and Ercetin[168], who conducted an observational study involving several advanced (as measured by TOEFL and IELTS scores) L2 speakers of English as they used an English-language hypertext system during think-aloud sessions. The system provided standard linguistic support features such as language glosses for certain words, as well as contextual annotations providing additional background information about the topic of the text. Subjects in this study made heavy and regular use of both features. The subjects were divided into two groups based on their level of background knowledge of the system's contents; both groups used the language support features, although the group with more background knowledge used them somewhat less than the less-knowledgeable group. One subject commented on the cognitive importance of annotations and glossary support when reading in an L2 language: "if the word is essential for comprehension and I cannot figure out the meaning, *I lose my motivation to read*" (emphasis ours). Perhaps due to the presence of easy access to glosses and annotations, the subjects relied on context to determine word meaning only infrequently. In this and other respects, Akyel and Ercetin's findings mirrored those by Chun[169].

Overall, the literature surrounding linguistic support tools for second-language readers is inconclusive, and is often hampered by sample size or other experimental design issues. However, the general findings appear to be that while users appreciate having such features, it is unclear as to whether they actually help them to read in a second language. Note that most of these sorts of studies focus on *information extraction* tasks rather than *search* tasks, and it may be that such tools would be more useful in a different context.

4.2 CLIR in Biomedicine

We will close this chapter with a brief discussion of cross-language information retrieval in biomedicine. As discussed in Chapter 3, biomedicine is an extremely information-rich discipline, and as shown in Chapter 2, language represents a significant barrier to many of the potential beneficiaries of this bounty of information. As such, CLIR tools have the potential to be very valuable to both researchers in and practitioners of biomedical professions. In general, medical CLIR has focused on providing multilingual access to MEDLINE content (as opposed to other forms of medical content), although there are exceptions, notably including Roseblat, et al.'s use of the PAHO medical machine translation system to enable multilingual access to the US government's ClinicalTrials.gov database of clinical trials[170].

The medical domain, like other specialized domains[171], has a sufficiently specialized vocabulary that many general-purpose machine translation systems exhibit suboptimal performance.[172]¹⁷ Studies by Zeng-Treitler, et al.[11] and Bedrick & Mauro[174] have found that existing commercial machine translation systems are inadequate for unsupervised use for clinical purposes. As such, most medical CLIR systems make use of additional knowledge sources to enrich their translation capabilities, typically the National Library of Medicine's Unified Medical Language System metathesaurus. However, other approaches have been used, for example, Daumke, et al.'s "Morpho-Semantic Indexing"[175]. Additionally, Lu, et al. have had a great deal of success at mining generic WWW text to develop Chinese-English MeSH mappings[176, 177].

One early system was SAPHIRE International[13], which enabled users to search across an early collection of clinically-oriented web sites using queries written in multiple languages as well as MEDLINE. It used a classical dictionary-based ap-

¹⁷ Of course, medical translation in general— even when done by humans— can be a difficult task.[173]

proach based on the aforementioned UMLS Metathesaurus. Although SAPHIRE itself is not currently operational, it does have a very prominent spiritual successor in the form of BabelMesh[178, 12], which allows users to build PubMed queries in a variety of languages. While the system's translation accuracy is limited and variable (depending as it does on MeSH dictionary mappings), it has the distinction of actually being designed for use by end-users, and has been received favorably by them.[12]

BabelMesh features an innovative query interface, which uses "autocomplete" to assist users in constructing their queries (see Figure 4.1). This feature, while potentially useful, highlights BabelMesh's dependency on the Medical Subject Headings terminology, as its autocomplete options are relatively limited. However, this architectural decision on the part of the BabelMesh developers has enabled it to be extended to additional languages, such as Arabic[179].

One shortcoming of the BabelMesh interface is that while the query formulation screen is relatively advanced, the results display screen (see Figure 4.2) is almost entirely in English. Obviously, users with weaker English proficiency might find this interface to be less than optimal. As noted by Oard[113], many users of CLIR systems will ultimately need to obtain full translation of at least some of their results; however, in a clinical context, this would at best represent a barrier to use (see chapter 3 for discussion on how clinicians' use of information resources is affected by barriers to use) and at worst would make the system unusable. This problem is hardly unique to BabelMesh; Figure 4.3 illustrates a Spanish-language search interface with a fully-translated interface... but with English-language results. It is unclear whether or not this linguistic discontinuity is problematic for users.

As machine translation becomes more widely-used in biomedicine, we would do well to remain cognizant of its imperfect and limited nature. We would like

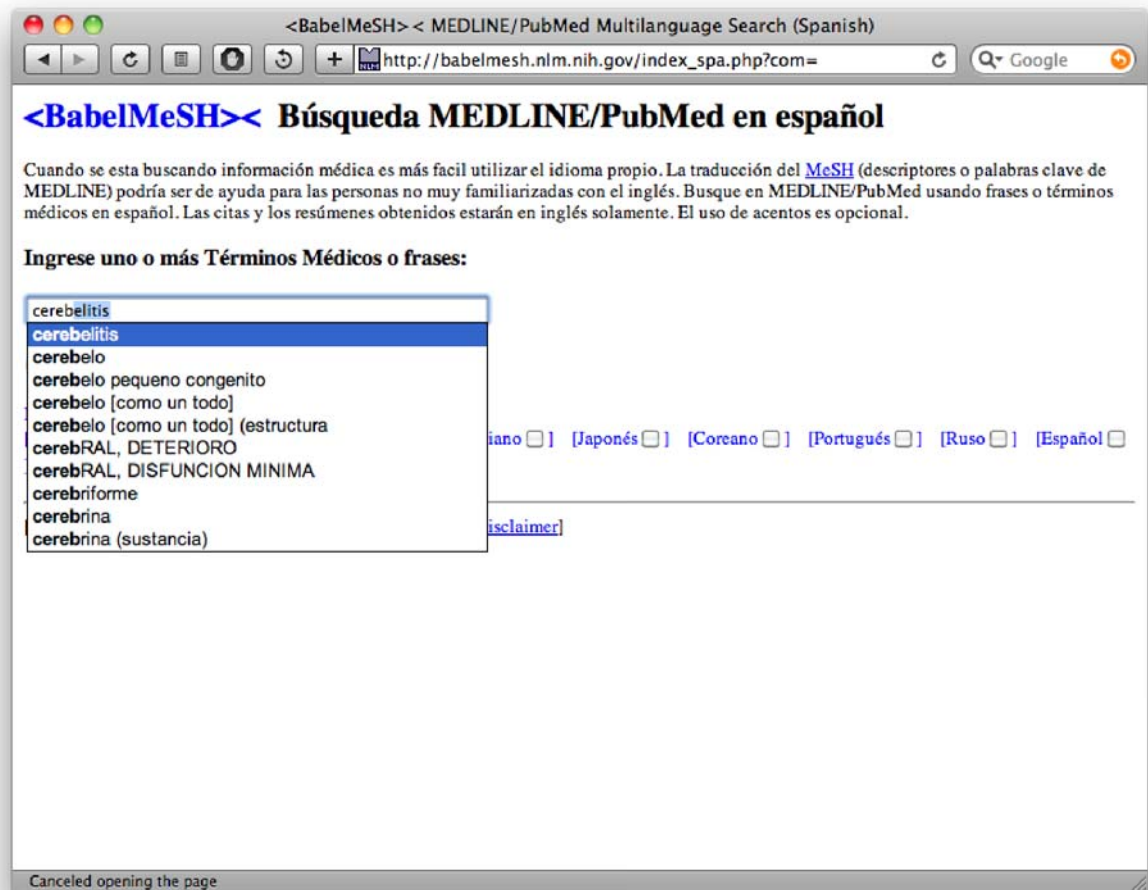


Figure 4.1: The BabelMesh query interface, including automatic query auto-completion. Note, however, that some of the autocomplete options are somewhat nonsensical (“cerebelo [como un todo]”), resulting from the system’s dependency on the contents of MeSH. Note further the completely internationalized interface (i.e., all UI elements are in Spanish, not English). Contrast to the results screen, shown in Figure 4.2.

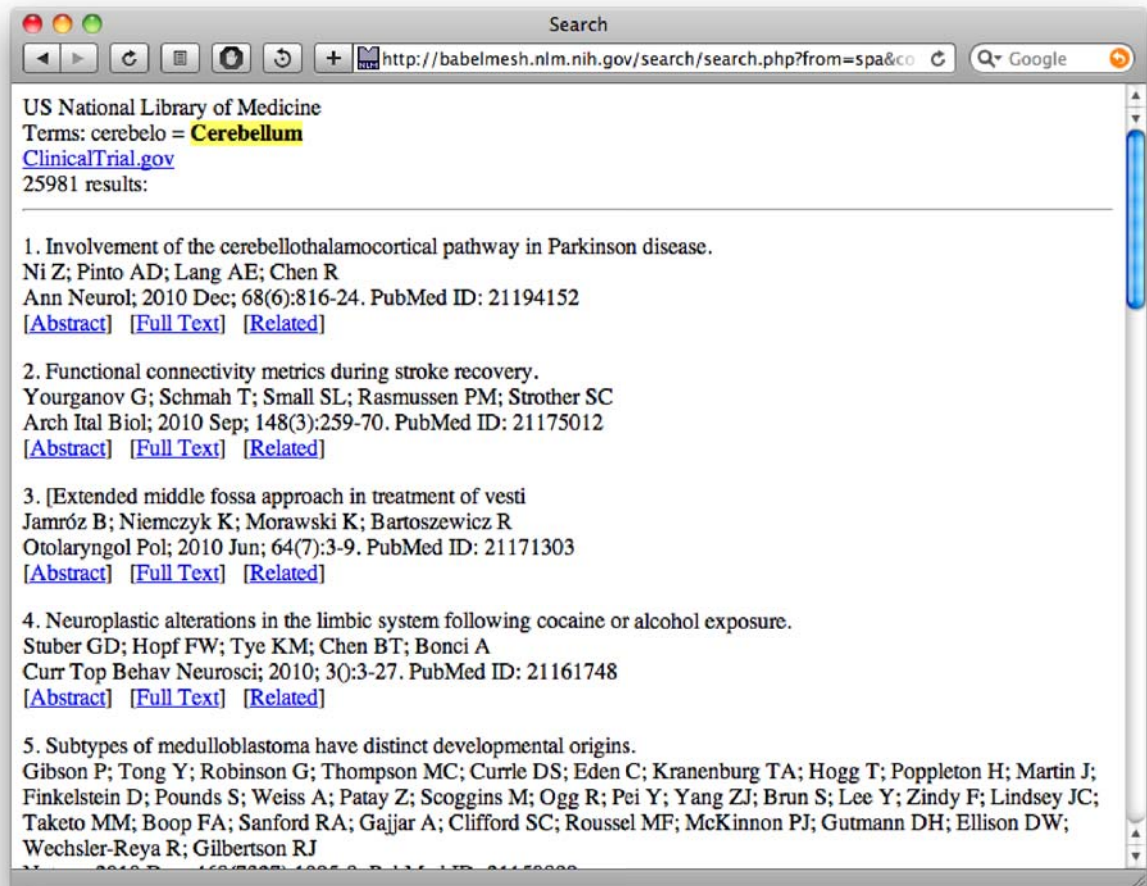


Figure 4.2: The BabelMesh results presentation interface. Note that it indicates the English equivalent to the user's query... but also note that all of the results are presented in the original English, and that the various user interface elements are also left untranslated ("full text," "abstract," etc.).

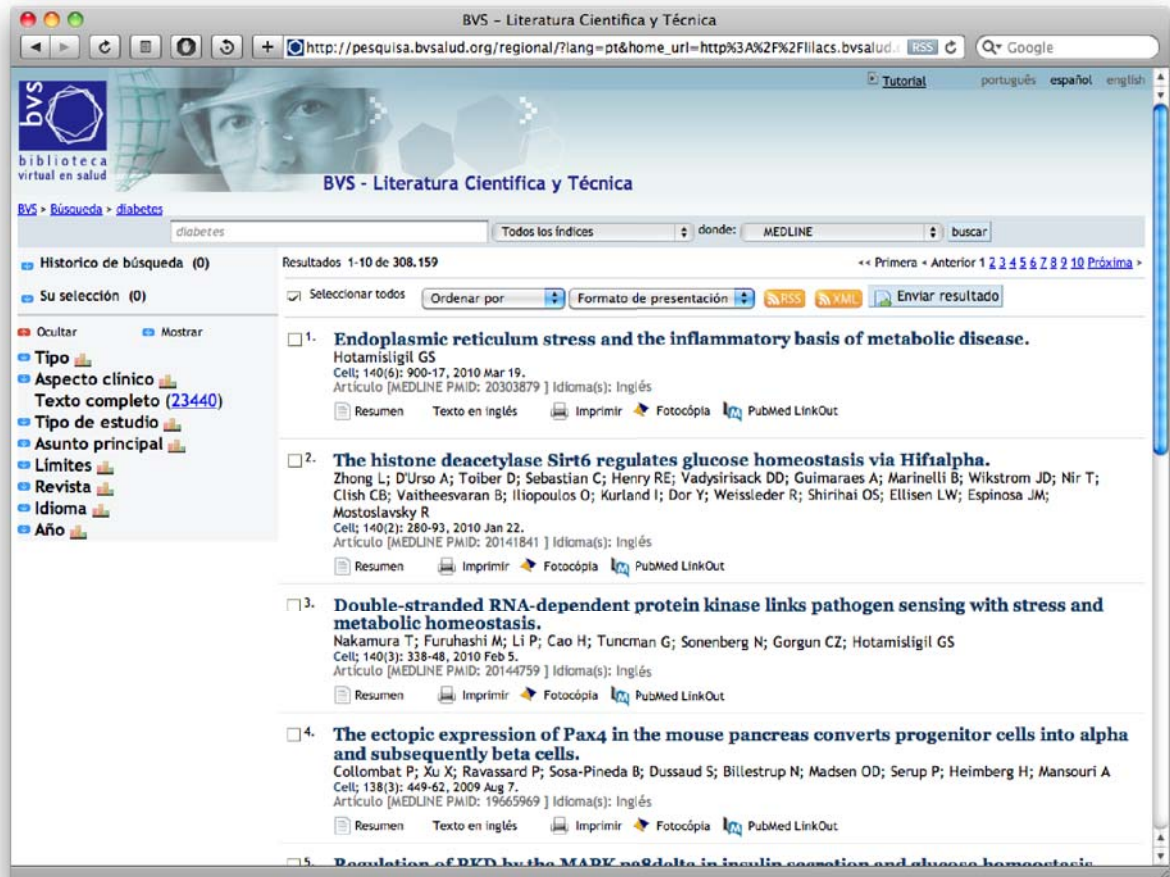


Figure 4.3: The Brazilian “Virtual Library of Health’s” Spanish-language search interface, displaying the results of a English-language MEDLINE search. Note the Spanish-language user interface elements (e.g., the “Aspecto clínico” facet browser, etc.) contrasted against the English-language content.

to conclude this chapter by relating a cautionary tale about a particular category of translation failure— one to which we feel machine translation is particularly susceptible.

It is common for non-linguists to assume that all words have one-to-one correspondences between languages. This cognitive fallacy— often known as the “naïve lexical hypothesis”[180]¹⁸— has been the source of countless mishaps, both large and small, throughout human history. In an unclassified 1968 article from the National Security Agency’s *Technical Journal*, an anonymous author[182] describes how this cognitive fallacy may be partially at fault for the atomic destruction of Hiroshima and Nagasaki at the end of the Second World War.

In July of 1945, the Allied forces issued an unconditional demand for Japanese surrender that became known as the “Potsdam Declaration.” The Japanese government’s response to the Allies’ demands, as communicated in the *Asahi Shinbun* newspaper’s account of the press conference, used the word “*mokusatsu*”, which in Japanese has several possible meanings. All of *mokusatsu*’s meanings relate to the concept of “silence,” but they range in tone from “contemptuous silence” to simply “withholding comment.” In fact, *mokusatsu* is apparently often used by Japanese politicians in much the same way as American politicians use the phrase “no comment.”

Unfortunately, for unknown reasons, the international news agencies reporting on the press conference did not mention this linguistic ambiguity. Instead, they simply stated that “in the eyes of the Japanese government the ultimatum was ‘not worthy of comment.’” The Allied commanders interpreted this as a sign of defiant resistance on the part of the Japanese government, and ten days later dropped an atomic bomb on Hiroshima.

Among the many lessons we may learn from this episode is that a translation

¹⁸ Weaver’s 1955 discussion of this topic is also particularly illuminating[181], and proved influential in the development of machine translation in the mid-twentieth century.

that allows the reader to remain blissfully unaware of potential ambiguity lurking beneath its surface is doing the reader a disservice. As the author of the NSA article states:

If a word is capable of variant translations, this fact should be conveyed to the reader... As a matter of principle, that unknown translator should have pointed out that [the] word has two meanings, thereby enabling others to decide on a suitable course of action.

However, most (if not all) machine translation systems fail to provide their users with such awareness of ambiguity in the translations they produce. In fact, MT systems typically present their users with a pleasant fiction: that their translations are neat, tidy, and accurate. Text goes into the system encoded in one language, and comes out the other end encoded in a different language. In reality, of course, machine translation— particularly *statistical* machine translation (SMT)— is often neither neat, tidy, nor accurate. Consider the first part of the methodology's name: "statistical". SMT is by its very nature imprecise; its translations are, by definition and by design, the probabilistic "best" that a given algorithm can do with a given set of training data.

Why, then, do the user interfaces of SMT systems persist in promoting the naïve lexical hypothesis? When a translation by a SMT system "succeeds" in producing a translation that meets a user's immediate need, the user is left with an inaccurate sense of certainty in the translation (and perhaps in how well the translation met their need). When the system "fails" (i.e., produces a translation which does *not* meet the user's immediate need), the user is left with no understanding of why, and feels frustrated by and alienated from the software.

Instead, why not build support for ambiguity into the user interface? Indicate to the user how and why the system ended up with the translation that it did, perhaps including alternative meanings for words, or even alternative translations

for entire passages together with confidence levels for each one. This could lead to more sophisticated users of SMT, which in turn could help prevent “*mokusatsu* moments” in the future.

This is particularly relevant to the medical domain, in which seemingly minor changes in wording can have serious consequences. It is incumbent on developers of systems that use machine translation to build their user interfaces in such a way as to empower their users without endangering them.

Chapter 5

Research Statement

As discussed in Chapter 3, clinicians routinely experience many different kinds of information need, and have a plethora of resources available with which they may attempt to meet those needs. However, they also face numerous challenges that prevent these resources from being as useful as they could be. For many of the world's clinicians, language represents an important example of such a challenge, as most internationally-published medical literature is in English. The previous chapter described a number of different technical approaches that system developers in other fields have used to address these challenges, along with various methods they have used to evaluate those approaches.

This dissertation's research goals are twofold. First, we seek to take some of what has been learned about interactive cross-language information retrieval and apply it to the medical domain. Specifically, we will use modern machine translation technology to provide MEDLINE search results to Spanish-speaking users in their native language by means of an integrated Spanish-language literature search system. We believe that many Spanish-speaking clinicians will find Spanish-language results easier to read than English-language results, and that, as a result of this, will be better able— and more likely— to make use of current

medical literature.

Our second research goal is to design and field-test a re-usable methodology for use in evaluating issues such as this one. As seen in Chapter 4, the field of information retrieval has a long history of evaluation, but relatively little of it has focused on issues of evaluating interactive systems. Our hope is that our methodology will help make future evaluations more standardized and repeatable, and that our protocol's metrics may lead to better comparability between studies.

Our evaluation's research questions are:

1. Do Spanish-speaking clinicians have discernible preferences regarding the language and manner in which search results are presented? In other words, do they generally prefer machine-translated Spanish-language results to English results, or do they prefer viewing results in their original language?
2. Does the language in which results are presented affect Spanish-speaking users' performance at using literature search systems?
3. What role does English proficiency play in an individual user's preferences or performance?

The remainder of this dissertation describes the implementation of our system, the design of our experimental methodology, and the results of a user study designed to investigate these research questions.

Chapter 6

The BuscTrad System

A distributed system is one in which the failure of a computer you didn't even know existed can render your own computer unusable.

Leslie Lamport[183]

The BuscTrad system is a multilingual, web-based information retrieval system that allows users to perform searches over biomedical literature, and to access the results of those searches in English, Spanish, or a combination of the two. Its *primary purpose* is to enable Spanish-speaking users to more easily find and select English-language scientific and medical articles. To accomplish this, the system uses machine translation (MT) to translate various parts of English-language search results (titles, abstract, etc.) into Spanish before presenting them to the user. The fundamental premise is that users will have an easier time identifying and reading relevant titles and abstracts if they are able to do so in their native language.

The system's *secondary purpose* is to serve as a platform for evaluations of user behavior and performance. As such, the system's architecture is highly modular, and supports extensive logging of user activity. It also features several use modes

designed to support its use in user studies. This enables investigator-users to easily experiment with different aspects of the system's behavior or appearance without affecting other parts of the system.

BuscTrad is written in the open-source Ruby programming language¹ using the Ruby on Rails (RoR) web framework². Released in 2004, RoR is an open-source Model-View-Controller framework whose foci are programmer productivity and sensible code organization. BuscTrad is currently operational and publicly available, and can be accessed at the following address:

<http://skynet.ohsu.edu/busctrad>

6.1 High-level architecture

The BuscTrad system's architecture and general information flow are described in Figure 6.1. The system is comprised of a pipeline of four major modules, each of which may easily be changed in an independent manner:

1. Query
2. Search
3. Translation
4. Results Presentation

In its present form, BuscTrad relies heavily on external web services for its functionality. A full discussion of what constitutes a "web service" is outside the scope of this dissertation; for our purposes, the definition used by the World Wide Web Consortium (W3C) will suffice: "a software system designed to support interoperable machine-to-machine interaction over a network."^[184] A web service is dis-

¹<http://www.ruby-lang.org>

²<http://rubyonrails.org/>

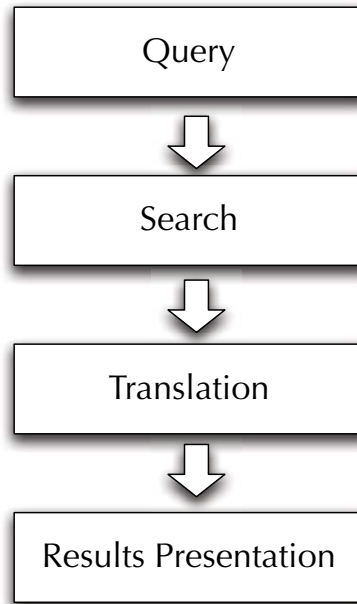


Figure 6.1: The major architectural modules of the BuscTrad system. Arrows indicate data flow between modules during system operation.

tinct from other approaches to network-based machine-machine interoperability (such as CORBA or DCOM) in two important respects. First and foremost, web services use the standard Hypertext Transfer Protocol (HTTP) to exchange data, as opposed to a custom network protocol. HTTP is the protocol that underlies all traffic on the World Wide Web (WWW), and is both simple and widely-implemented. The second important distinguishing characteristic of web services is that the data exchanged by web service applications are typically in the form of human-readable messages encoded in XML or some other textual format, as opposed to being encoded in an opaque binary format.

While a full discussion of HTTP is beyond the scope of this chapter,³ some knowledge of its mechanics will be helpful to the reader, as numerous aspects of BuscTrad (and of the evaluation management system described in Chapter 7) depend on or were constrained by HTTP and its functionality. The hypertext transfer

³ See Internet Engineering Task Force RFC #2616 at <http://www.ietf.org/rfc/rfc2616.txt> for the complete HTTP specification.

protocol is a stateless client-server network protocol, meaning that each HTTP interaction takes place between two computers, and that each interaction occurs entirely independently of any other interactions that may have occurred in the past or may be occurring simultaneously (i.e., the protocol does not maintain state between interactions). The computer that initiates the interaction is called the *client*, and the computer with which the client interacts is called the *server*.

An HTTP interaction begins when an HTTP client submits a *request* to an HTTP server, which then processes the request and sends a *response* back to the client. The protocol was designed to facilitate the transmission of files and documents from one computer to another, and so in a very real sense one may think of the protocol as a sort of primitive (but very formal and strict) language that computers may use to request files from one another, and to respond to those requests.

An HTTP request consists of one of a small set of “verbs” (“GET”, “POST”, “DELETE”, etc.), and a target resource (typically a file or other URL) that the verb is meant to “act” upon.⁴ So, for example, a request reading “GET /about_us.html” would represent an instruction to the server to retrieve a file named “about_us.html” located at the top (“root”) of the server’s file hierarchy. The server’s response may contain any kind of data, but typically consists of the contents of a document authored in HTML or an image. In the case of this example GET request, the response would consist of the contents of “about_us.html”, whatever those might be. Whatever is sent down in the response is called the “body” or “payload” of the response.

In addition to these common elements (method, resource, etc.), valid HTTP requests and responses always begin with in-band control metadata (known as “headers”). These headers are key-value pairs that provide additional instructions to the server (in the case of request headers) or the client (in the case of response headers). For example, a typical request will include headers specifying the ver-

⁴ The “verb” used in any given HTTP request is formally referred to as the request’s “method.”

sion of the HTTP protocol that the client wishes to use, a list of the languages in which the client is willing to receive its response, and so on. A typical HTTP response might contain headers specifying the file type of the response (e.g., whether it contains a JPEG image, a Microsoft Word document, etc.), and whether the server expects the client to cache the response.

The most important response header is a code indicating whether the client's request was successfully processed by the server, and, if not, the reason why. The HTTP standard specifies error codes for a variety of situations: if the client requests a non-existent file; if the request itself was malformed; if the server encounters an internal error while processing the request; if the file requested by the client has moved to a different location; and so forth. HTTP clients generally are able to handle some types of errors better than others. For example, if an HTTP server responds to a request with an error indicating that the requested file has moved to a new location, most modern HTTP clients (including nearly all web browsers) will automatically prepare a new request aimed at that new location without any interaction on the part of the user. This is known as an "HTTP Redirect."

The protocol was originally designed with file retrieval in mind, and, therefore, early HTTP servers could do little more than respond to requests for specific files—they could only handle requests for resources that happened to be files on the server's filesystem. However, programmers quickly realized that they could enable richer functionality if they extended the HTTP servers' functionality by connecting them to external programs. Typically, this is done by configuring the server to respond to requests for certain URLs by executing a specific program and using its output as the body of the HTTP response to the client (instead of sending a particular file's contents, as they normally would).

Compared to many other network protocols, HTTP is very straightforward, and may easily be used as the foundation for many different types of applications.

Most HTTP server software packages make it trivial to expose external programs or scripts via a URL, and over the years a vast ecosystem of specialized programming languages, tools, and libraries has developed to better support programmers in this task.

Due to HTTP's simplicity and ease of integration, as well as the ubiquity of the World Wide Web in modern telecommunications, building a web service is a relatively easy and straightforward way for programmers to provide software to their users. There are many examples of biomedical applications making use of web services; for example, see [185, 174, 186], etc. Many organizations and companies have used web services to expose various aspects of their functionality in a machine-consumable way. For example, the National Library of Medicine provides a web service that allows third-party programs to access content from MEDLINE and other NLM databases, and both Google and Yahoo provide web service interfaces to their search engines. These sorts of interfaces are often grouped together with other application programming interfaces (APIs), and are represent an increasingly important tool for linking together computing resources from different organizations.

6.1.1 Caching

One important factor that must be taken into consideration when using distributed resources such as these is that of performance: a system that relies on third-party resources (e.g., external APIs) is, in effect, held hostage by the slowest of the resources on which it depends. The problem is exacerbated further by the fact users of web pages are notoriously sensitive to delays in page load times[187, 188, 189].⁵

In order to help minimize BuscTrad's page load times, the system automatically

⁵For our purposes, "page load time" refers to the amount of time that elapses between the time at which the user initially requests a specific URL and the time at which they are able to interact with it.

and transparently caches any remotely-retrieved data, including search results, translations, etc. using Memcached,⁶ an open-source in-memory object caching system (sections 6.3 and 6.4 discuss specifics of how BuscTrad makes use of Memcached). Object caching is a technique wherein frequently-used pieces of data that would ordinarily need to be retrieved from a slow or unreliable source (i.e., from a remote network service, or a complex database query) are instead stored in the computer's main memory, which is one of the fastest places it can be stored—thereby making future requests for the same piece of data orders of magnitude faster.

Memcached is a particularly popular object caching system among developers of large-scale web sites, in part due to its high performance and reliability, its ease of administration, and its ability to distribute itself among an arbitrary number of servers. This means that a single Memcached installation may make use of memory on multiple connected computers, which in turns means that Memcached may be quickly and easily expanded should more caching capacity become necessary.

Along with most object caching systems, Memcached is quite simple to use from the perspective of a software developer. Essentially, the programmer may view a Memcache cache as a large key-value store, similar to a Perl dictionary or a classical hash table. Checking to see whether the cache contains a value for a given key is extremely fast (so fast as to be, for all intents and purposes, free), as is retrieving or storing an arbitrary object. In Memcached, the key is an ASCII or Unicode string, and the value is either another string or a piece of binary data. Other object caching systems offer support for more highly structured keys or values; Memcached's focus is instead on simplicity and ease of use.

Programmers do not typically interact directly with Memcached; rather, they typically use a library of some sort. In many languages, these libraries provide

⁶<http://memcached.org/>

additional features. For example, the Memcached library that BuscTrad uses is written in the Ruby programming language, and supports automatic serialization and deserialization of Ruby objects into strings so that they may be stored transparently in a Memcached cache.

The following sections discuss each of BuscTrad's major architectural components (as illustrated in Figure 6.1) in turn.

6.2 Query

The query module has two responsibilities: collecting a raw query from the end-user, and doing any necessary post-processing of that query (term expansion, stop-word removal, etc.). Since our focus thus far has been on experimenting with different result presentation modes, the query module's user interface is relatively simple. However, there are many user interface modifications that we would like to experiment with in the future (see [71] for a thorough overview of the range of user interface approaches that others have tried).

As mentioned, the Query module is also responsible for post-processing of user queries before handing them off to the Search module. For the purposes of this dissertation, the Query module does little to no post-processing, as the Search module (see Section 6.3) requires that user queries be passed through verbatim. However, if BuscTrad were at some point to be extended so as to perform translation of user queries (from Spanish to English, for example) the Query module is where that translation would take place.

6.3 Search

The version of BuscTrad described herein uses the National Library of Medicine's "Entrez" application programming interface (API)[190] to execute the queries received from the Query module against the PubMed database. The Entrez API allows third-party developers to access much of the same functionality as the PubMed website itself. For example, developers may use the Entrez APIs to issue MEDLINE queries using PubMed query syntax, retrieve information about specific MEDLINE entries, search for full-text resources and article metadata from PubMed Central, and so on. Other search modules would search different databases (e.g. GenBank), combine results from multiple sources, or otherwise perform different searches.

As mentioned in Section 6.2, the Query module currently does little to no post-processing of user queries. This is because the specific APIs that the Search module uses do the same sophisticated query processing as the PubMed website: stop-word removal, boolean term grouping, mapping user query terms to MeSH headings, etc.

At the implementation level, the search functionality is encapsulated in such a way as to easily enable other repositories (besides PubMed) to be used in BuscTrad. The actual communication with Entrez is done using the open-source BioRuby library[191]. As discussed above, BuscTrad's repository interface caches its search results using Memcached, which enables the site to minimize both the number of calls it must make to the Entrez APIs as well as the amount of data that each call must transfer, which in turns help minimize page load times as perceived by the user.

Retrieving PubMed entries using Entrez is a two-step process. First, the client application (in this case, BuscTrad's `RepositoryInterface` class) submits its query by making a call to the Entrez API's "ESearch" method. This returns an or-

dered list of PubMed IDs (PMIDs), which represent the results of the search query. Then, the client application must make a second API call, this time to the “EFetch” method. EFetch takes a list of PMIDs and retrieves their corresponding MEDLINE records. While there are limits to how many PMIDS may be retrieved at once using EFetch, they are high enough that BuscTrad almost never needs to make multiple EFetch calls during the course of a given search. This is an important feature of the Entrez APIs— otherwise, if a call to ESearch retrieved 50 PMIDs, we would need to make 50 calls to EFetch in order to retrieve all of the results, with predictably negative effects on our page load time.

BuscTrad’s caching behavior comes into play between the call to ESearch and the call(s) to EFetch. After retrieving a list of PMIDs that satisfy the user’s query using ESearch, `RepositoryInterface` checks BuscTrad’s Memcached cache for each PMID in order to see whether its corresponding complete MEDLINE record object has previously been retrieved and stored. Any such MEDLINE records are set aside, and the remaining PMIDs are passed off to EFetch. After retrieving the results, the `RepositoryInterface` class uses BioRuby’s library routines to parse MEDLINE records into Ruby objects, and then combines the two lists of results (one comprised of cached results, the other containing the new results) into a single collection of results, sorted in the order that they had been in originally in the response from ESearch. As a final step, the fresh results are themselves stored in Memcached for future retrieval (see Figure 6.2 for a graphical representation of this process).

This is so that, should any subsequent queries retrieve these particular records, they will not need to be retrieved (relatively slowly) via EFetch and will instead be found (almost instantaneously) in the cache. While, in this case, caching does not reduce the number of calls that BuscTrad must make to external services, it *does* reduce the amount of data that must be transferred during each call, and also reduces

the amount of work that the external service must perform in order to respond to BuscTrad's requests, which together result in faster response and processing times.

This speed improvement has significant usability benefits. Modern information-seeking is an interactive process, and it is normal for a user to iteratively refine their query by repeating and varying it multiple times. When carrying out this process with a search engine (e.g., Google, BuscTrad, or PubMed), it is not uncommon for variations on a given query to retrieve overlapping result sets. When this occurs with BuscTrad (i.e., when a new search retrieves some of the same results as were retrieved by the a previous search), the system is able to return its results to the user much more quickly than it can when a search returns completely novel results.

6.4 Translation

The Translation module is responsible for translating the results from the Search module into the target language⁷ and delivering them to the Results Presentation module. For the purposes of this dissertation, the Translation module makes use of the Google Translate[192] machine translation (MT) system to translate English-language search results into Spanish-language results. However, because of the modular design of the entire system, it would be relatively simple to build an alternative module that used a different MT system.

6.4.1 Why Google Translate?

There are three reasons why the current BuscTrad translation module uses Google Translate (as opposed to a different machine translation engine):

⁷In cross-language information retrieval, the terms *source language* and *target language* are used to refer to the languages in which a system's corpus of documents are written in and that which the system is translating into, respectively. In the case of this particular CLIR system, the *source* language is English, and the *target* language is Spanish.

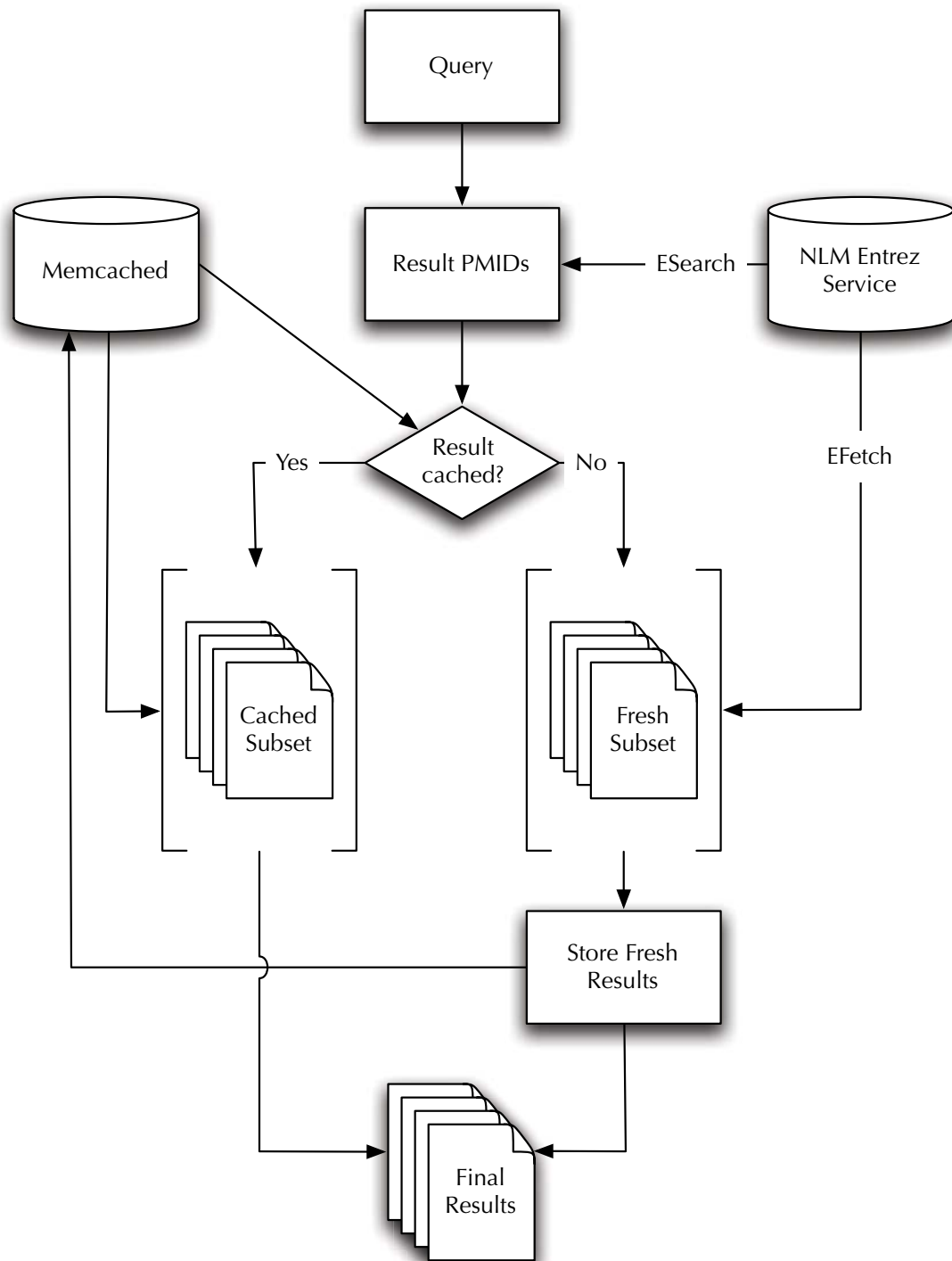


Figure 6.2: The flow of data within BuscTrad's search module. Arrows indicate the flow of data between the various steps and data sources.

1. *Cost*: Google Translate is freely available for use;
2. *Technical Convenience*: Just as the NLM provides public APIs to enable third-party developers to access PubMed, Google provides a set of APIs to its translation system. In addition to simply translating text from one language to another, the Google Translate APIs also provide language detection and transliteration functionality, although the Translation module does not currently use these additional features.
3. *Widespread Use*: In our discussions with Latin American clinicians, we have learned that many of our target users are already making use of Google Translate, suggesting that its functionality is adequate for our users' needs.⁸

Of course, there exist a great many machine translation systems, and comparing their relative performance is a potentially fruitful topic for future work using the evaluation framework and modular architecture described in this dissertation (see chapter 11).

6.4.2 Implementation Details

Google provides two ways for users to interact with its translation service: through the public-facing website, <http://translate.google.com>, and through its APIs (described in detail at <http://code.google.com/apis/ajaxlanguage/>). The public-facing website allows visitors to enter source-language text and view that text's translation into any one of dozens of target languages (see Figure 6.3). The APIs are intended for use by third-party developers who wish to make use of Google's machine translation technology within their own applications, whereas the public-facing website is intended for use by human users.

⁸ In addition to this common-sense argument, there is evidence to suggest that user choice can sometimes be used as a proxy for more objective measures of system performance[163].



Figure 6.3: The public-facing Google Translate interface. Note that the translation system is able to handle non-Latin character systems. Note further that out-of-vocabulary words (“noisiest,” in this case) are left untranslated, but are otherwise left untouched by the user interface.

Initially, BuscTrad's interface to Google relied on the public API system. These APIs follow a common pattern among modern web services: to translate a string, a developer issues an HTTP "GET" request to a particular URL that includes encoded parameters such as the string's source language, the desired target language, and the actual text to be translated. The response to the request contains the search results, encoded as a data structure in the standard JavaScript Object Notation serialization format (JSON)⁹. JSON is commonly used for this purpose in web services due to its simplicity and ease of parsing.

While this approach worked well at first, we quickly discovered some important and poorly-documented limitations to Google's publicly-available APIs. The most important limitation related to the amount of text that could be translated in any single request. Recall that the public API operates by requesting a single URL containing encoded parameters, *including the string to be translated*. This means that, as the amount of text that one wishes to translate increases, so does the length of this URL. While there is no specified limit to how long a URL may be, for reasons of security and practicality, most web servers refuse to process requests for URLs that are longer than one or two thousand characters, and Google's server is no exception. Therefore, the amount of text that may be translated via the public API is limited by the maximum URL length supported by Google's servers (generally approximately 1,000 characters).

Since a typical MEDLINE abstract is more than 1,000 characters in length, we were forced to find a workaround. The first possibility we considered was simply splitting each abstract into smaller pieces, using the API to translate each piece, and then recomposing the abstract. While this would be technically quite simple, we were wary of using an approach that would involve an unpredictable number of requests. This would remove our ability to accurately predict the amount

⁹<http://www.json.org>

of time our system would require to service each incoming request; furthermore, this approach would produce more work for Google's servers than would a single-request approach. Although our application is unlikely to ever produce a noticeable amount of traffic from Google's perspective, it is generally considered good form among web developers to minimize the number of requests one sends to a remote service. Additionally, while currently Google Translate appears to operate at the sentence level, we are unsure of how important a sentence's surrounding context is to the system's performance, and were concerned that translating smaller sub-passages of text would affect our translation quality. For these reasons, we wished to find a way to enable longer passages of text to be translated within a single request.

The main Google Translate website allows users to translate blocks of text that are much longer than 1,000 characters. By analyzing the website's code, and by observing the network traffic generated during a Google Translate session that involved translating more than 1,000 characters, we were able to reverse-engineer the extended translation functionality of the main Google Translate website. The reverse-engineered protocol (hereinafter referred to as the "full" API, as opposed to the publicly-available API) is quite similar to the public one, with a few small (but important) differences.

First and foremost, the full API is available from a different URL than the public API. Secondly, requests via the full API must be in the form of HTTP "POST" requests as opposed to the "GET" requests required by the public API. Finally, the JSON returned by the full protocol is formatted slightly differently from that returned by the public API. We were able to extend our client library such that it used the public API for shorter blocks of text, and the full API for longer blocks of text.

While this approach does not have the shortcomings of the previous solution

(splitting each abstract into several smaller passages), it is not without its own drawbacks, the most important of which is its unofficial and unsupported nature. Since it relies on undocumented aspects of the Google Translate website's architecture (aspects that the authors of Google Translate undoubtedly did not intend to be used by third-parties), it is extremely fragile. Small changes by Google can completely break the ability of our system to translate longer passages of text, and such changes are quite common, as the Google Translate website is under constant development.

When these changes do occur, however, it is usually a very simple matter to restore service, as site changes are almost always cosmetic or syntactic in nature. To ensure fast detection of problems relating to changes on Google's part, and to facilitate quick repairs to such problems, we have developed and continue to maintain an extensive suite of automated tests for all aspects of our Google Translate interface. By running these tests, we can quickly see if any parts of the system are broken, and pinpoint precisely where such breakages are and what is causing them.

The Translation module also makes use of a caching layer in much the same way as the Search module discussed in the previous section: when asked to translate a string from language "A" to language "B," the Translate module will by default first check the system-wide Memcached store to determine whether a cached copy exists of the string's translation for that language pair. If so, it will simply retrieve the cached copy; if not, it will obtain the translation from Google in the manner previously described in Section 6.3, and then store the translated string in the cache (see Figure 6.4).

This process helps to minimize calls to Google Translate. Consider the case wherein a user performs a search, and then refines that search by including a similar but different set of keywords in the query. The two searches' results are likely to

overlap to a certain degree, and because the first search's translations were cached, the system will not need to translate the overlapping articles' titles and abstracts. In practice, this results in a much more responsive-feeling user interface, at the cost of some additional code complexity.

6.5 Results Presentation

The Results Presentation module controls how the output from the Search and Translation modules are presented to the end-user. BuscTrad currently has three different Results Presentation modes:

1. *Monolingual English*: The Monolingual English mode does not use the output of the Translation module, and simply displays the raw English output from the Search module (see Figure 6.5). All user interface elements on the page (navigation links, etc.) are in English.
2. *Monolingual Spanish*: The Monolingual Spanish mode makes use of the output from the Translation module, and displays only results that have been machine-translated into Spanish (see Figure 6.6). All user interface elements are in Spanish.
3. *Bilingual*: The Bilingual module displays the original English results alongside the translated Spanish-language results (see Figure 6.7). User interface elements are in Spanish; the only English on the page is the English-language results section.

Under the current design, all three modules share common user interface elements. In each, the user is presented with a list of search results. Initially, each element in the list is an article title, author list, and journal citation. Alongside each result are several links. The first, when clicked, displays its article's abstract

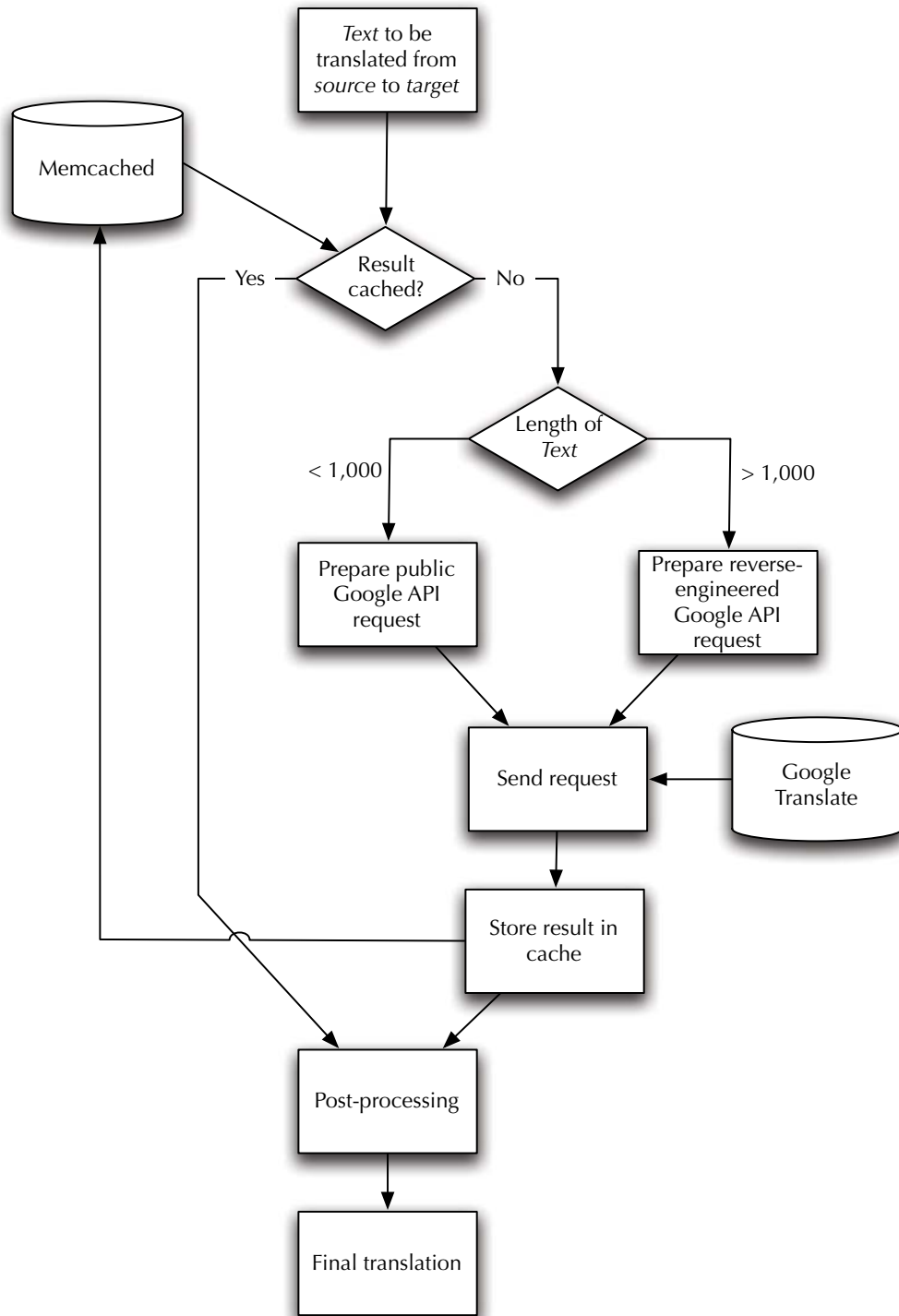


Figure 6.4: The flow of data within BusTrad's translation module. Arrows indicate the flow of data between the various steps and data sources.

(or hides the abstract if it is currently showing).¹⁰ Similarly, the second link toggles the display of a list of the MeSH headings assigned to the article (see Figure 6.8). If an article lacks an abstract or MeSH headings, the corresponding links are not shown. The third link opens its article's entry in PubMed in a new browser window, and the fourth link is used to "mark" or "select" that result for later perusal or download.

The MeSH heading display functionality is another place where BuscTrad makes use of external web services. In order to build the translated list of MeSH terms for each article, the system makes use of a terminology server that resolves English MeSH headings to their Spanish equivalents (using the Unified Medical Language System). If BuscTrad encounters a MeSH term without a Spanish major-heading equivalent, or in the case of MeSH qualifiers (which are not included in the Spanish-language version of MeSH), it falls back to using the Translation module as previously described.

In addition to these features, BuscTrad performs "query term highlighting," a commonly-used feature in which a search engine highlights occurrences of terms from the user's query when they appear in a result, often by changing the visual appearance of the text (e.g., printing occurrences of search terms in boldface, changing the background color, etc.) [71, 68]. The purpose of this is to help draw the user's eye to the passages within their search results that are most likely to be relevant to their query. In the case of BuscTrad, the system highlights query terms with a yellow background.

One important feature of BuscTrad's query term highlighting is that it takes place in both the English-language and the translated result listings (see Figure 6.9).

¹⁰There are two reasons for initially hiding the article abstracts. First, hiding the abstracts reduces visual clutter on the page, making it easier for users to browse through their results. Second, hiding the abstracts enables the system to "know" which articles the users find potentially interesting: to view the abstract, the user must actively request it from the system by clicking on a link. This activity may be logged by the system and analyzed later by investigator-users.

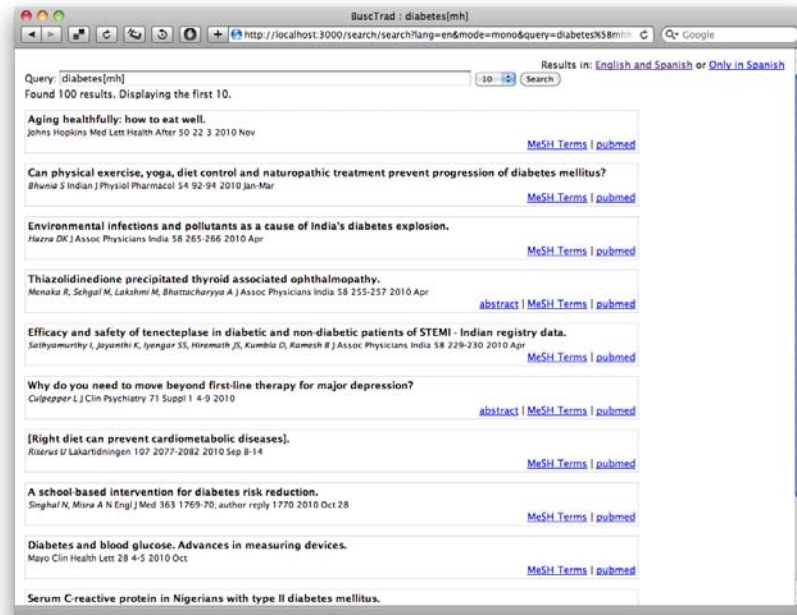


Figure 6.5: The Monolingual English results presentation module displays raw English search results (generated from the Search module; see Section 6.3) in an English-language user interface.

So, for example, if the user's query included the word "schistosomiasis," the system would not only print all occurrences of that word with a yellow background, but would also print any instances of "equistosomiasis" (its Spanish translation) with a yellow background, as well. The system uses the Translation module to determine which terms to highlight, and performs basic stemming (in both the original and translated text) in order to accommodate lexical variation between query and result terms (e.g., both "monkey" and "monkeys" are highlighted). While this relatively simplistic approach works tolerably well, this remains an area of active development, and there is certainly much room for improvement.

6.6 Pre-loaded result sets

Recall that the BuscTrad system has two purposes: first, to be search tool aimed at helping Spanish-speaking clinicians more easily access MEDLINE content; and,



Figure 6.6: The Monolingual Spanish results presentation module displays machine-translated Spanish-language search results (generated by the Translation module, see Section 6.4) in a Spanish-language user interface.

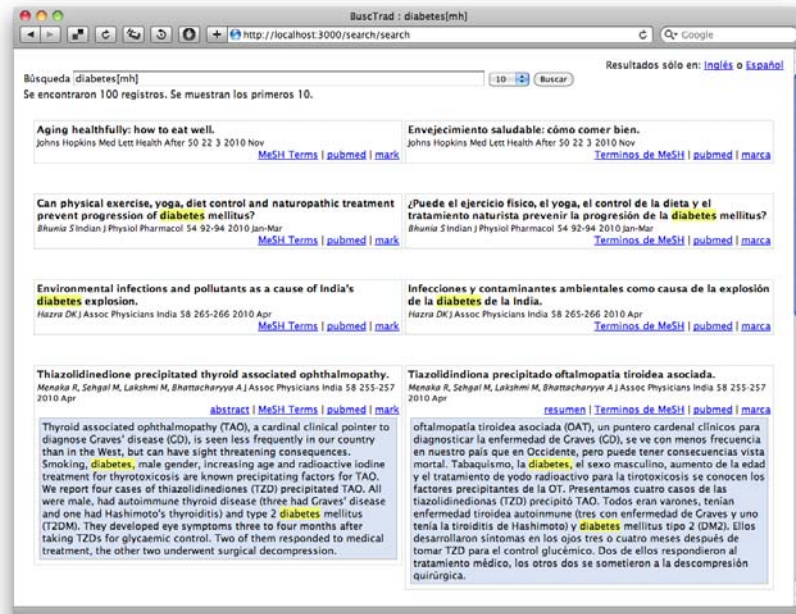


Figure 6.7: The Bilingual results presentation module displays untranslated results alongside their translated counterparts.

¿Puede el ejercicio físico, el yoga, el control de la dieta y el tratamiento naturista prevenir la progresión de la **diabetes mellitus?**
Bhunja S Indian J Physiol Pharmacol 54 92-94 2010 Jan-Mar
[Terminos de MeSH](#) | [pubmed](#) | [marca](#)

- Adulto
- *Diabetes Mellitus/*la terapia*
- **Dieta*
- Progresión de la Enfermedad
- **Ejercicio*
- Humanos
- Masculino
- **Naturopatía*
- **Yoga*

Figure 6.8: BuscTrad displays each result's MeSH headings, if present, including qualifiers. When possible, each English-language MeSH term is matched to its corresponding Spanish-language MeSH term using the Unified Medical Language System (UMLS). When there is no corresponding main heading, or in the case of qualifiers, which are not included in the Spanish-language MeSH release, BuscTrad uses its Translation module (see Section 6.4) to obtain an appropriate term. Headings and subheadings that the NLM indexer designated to be "major topics" are indicated in italics.

<p>Cationic nano-copolymers mediated IKKbeta targeting siRNA to modulate wound healing in a monkey model of glaucoma filtration surgery. <i>Ye H, Qian Y, Lin M, Duan Y, Sun X, Zhuo Y, Ge J Mol Vis 16 2502-2510 2010</i> abstract pubmed mark</p> <p>PURPOSE: To investigate the efficacy and safety of cationic nano-copolymers CS-g-(PEI-b-mPEG) mediated IkappaB kinase beta (IKKbeta) targeting siRNA in modulating wound healing in a monkey model of glaucoma filtration surgery. METHODS: The IKKbeta targeting siRNAs were chemically synthesized and screened in cultured monkey Tenon's fibroblasts in vitro. Fourteen monkeys underwent trabeculectomy and were randomly allocated to one of three treatment regimens: subconjunctival injection of either CS-g-(PEI-b-mPEG)/IKKbeta-siRNA (six eyes, 50nM, at the time of surgery and 7 days post surgery) or phosphate buffered saline (four eyes), or treated with mitomycin C (MMC; four eyes, 0.2 mg/ml). Bleb survival and characteristics, and intraocular pressure, were evaluated over a 60-day period. Histology of the surgical eyes was performed to evaluate ocular scarring and fibrosis in each group. RESULTS: Subconjunctival injection of CS-g-(PEI-b-mPEG)/IKKbeta-siRNA was well tolerated in this model. Both siRNA and MMC significantly prolonged bleb survival compared with the PBS group (the medians for survival days were 45.5, 60, and 29.5 in the siRNA, MMC, and PBS groups, respectively, p<0.01). Higher blebs were observed in the siRNA group than in the PBS group (p<0.01), while the MMC group showed the highest blebs among three groups (p<0.01). The surgical</p>	<p>Nano-copolímeros catiónicos mediada por siRNA IKKbeta dirigidos a modular la cicatrización de heridas en un modelo de mono de la cirugía de filtración de glaucoma. <i>Ye H, Qian Y, Lin M, Duan Y, Sun X, Zhuo Y, Ge J Mol Vis 16 2502-2510 2010</i> resumen pubmed marca</p> <p>PROPÓSITO: Para investigar la eficacia y la seguridad de copolímeros catiónicos nano-CS-g-(PEI-b-MPEG) mediada IkappaB quinasa beta (IKKbeta) la focalización del siRNA en la modulación de la cicatrización de heridas en un modelo de mono de la cirugía de filtración de glaucoma. MÉTODOS: El objetivo IKKbeta siRNAs fueron sintetizados químicamente y proyectado en el mono cultivo de Tenon fibroblastos in vitro. Catorce monos sometidos a trabeculectomía y fueron asignados al azar a uno de tres regímenes de tratamiento: la inyección subconjuntival de cualquiera de CS-g-(PEI-b-MPEG) / IKKbeta-siRNA (seis ojos, 50nm, en el momento de la cirugía y cirugía 7 días después) o tampón fosfato salino (cuatro ojos), o tratados con mitomicina C (MMC, cuatro ojos, 0,2 mg / ml). la supervivencia de la ampolla y características, y la presión intraocular, fueron evaluados durante un periodo de 60 días. Histología de los ojos quirúrgica se realizó para evaluar la cicatrización ocular y fibrosis en cada grupo. RESULTADOS: la inyección subconjuntival de la CS-g-(PEI-b-MPEG) / IKKbeta-siRNA fue bien tolerado en este modelo. Ambos siRNA y MMC significativamente la supervivencia prolongada ampolla en comparación con el grupo PBS (la mediana de los días de supervivencia fueron del 45,5, 60, y 29,5 en los grupos de siRNA, MMC, y PBS, respectivamente, p <0,01). ampollas</p>
--	---

Figure 6.9: BuscTrad identifies terms in search results that match terms found in the user's query ("monkeys," in the case of this example query), and highlight them in yellow in order to draw the user's eye to passages that are presumably more likely to be relevant. Note that the highlighting works across languages, and also takes place on lexical variants of query terms (singular vs. plural, etc.) in both languages.

second, to serve as a experimental platform for evaluating various aspects of information retrieval user behavior. In the service of that second purpose, BuscTrad has the capability to display pre-loaded search results to users using any of the Result Presentation modes (English-only, Bilingual, etc.) in addition to its previously-described MEDLINE-searching capabilities. This means that investigators wishing to maintain a consistent experience for their subjects may pre-load the system with any number of article sets, which the system may then present to subjects as though they were natural search results. This aspect of BuscTrad is entirely separate from the rest of BuscTrad except insofar as it makes use of the result presentation modules described in Section 6.5. It is therefore not included in the system diagram in Figure 6.1, as it is not intended for use by end-users but rather by researchers designing studies involving BuscTrad.

The system is capable of automatically loading article sets via PubMed query (see Figure 6.10). The process is simple. First, the investigator enters a PubMed query whose results contain the desired article set, after which BuscTrad runs the query and saves the results for future use (see Figure 6.11). BuscTrad can then be configured by the investigator to display the contents of the new article set using any of the previously-discussed presentation modes (Figure 6.12).

As new article sets are loaded into BuscTrad's database, the system can optionally load translations for those articles using the previously-described Translation module (see Section 6.4), and can save that translated content in a persistent manner. This is important for studies making use of such translations, as Google Translate's performance is constantly changing— meaning that there is no guarantee that the same text will translate the same way twice. This would represent an obvious source of bias in any experiment that depended on translations of pre-loaded articles, in that two subjects who viewed the same article on different days might potentially see a different translation. The problem is compounded by the fact that

Google Translate's changes generally represent what are meant to be *improvements* in translation quality, such that our second subject might be viewing a significantly better translation than our first subject (or vice-versa, should the changes turn out to be regressions rather than improvements in translation quality). To control for this, the system has the capability to pre-fetch and store translations of all articles loaded using this set of features.

In addition to loading new article sets based on arbitrary PubMed queries, BuscTrad can perform various operations on article sets already in its database. For example, the system is able create new article sets by splitting existing article sets into randomly-composed partitions, randomly subset any number of articles from an existing article set into a new article set, or merge two existing article sets together with user-specifiable proportions. For example, given two existing article sets *A* and *B*, a user could create a new article set *C*, 25% of which was originally from set *A* and 75% of which was from set *B*. Figure 6.13 illustrates this process.

Note that entries in article sets maintain a reference to their "parents." So, in the example shown in Figure 6.13, individual entries in the "Monkeys & Bilharzia" set will "know" whether they originated in the "Monkeys" or "Schistosomiasis" sets. This is important, as it allows us to use each article's origin as a "label" for use in calculating various statistics. For example, we might design an experiment involving the "Monkeys & Bilharzia" set such that the articles originating from the "Monkeys" set were "relevant" (for the purposes of calculating precision, recall, etc.) and those originating from the "Schistosomiasis" set as "not relevant."

Section 7.6.1 describes one possible use scenario for pre-loaded result sets. BuscTrad includes slightly-modified versions of the standard result presentation modes designed to work with pre-loaded result sets, and any researcher wishing to design their own custom result presentation mode can do so using standard HTML design tools.

New from pubmed query:

Title:

Query:

Figure 6.10: Article sets may be loaded into the system by investigators by specifying a PubMed query, whose results will be fetched and then saved for future use.

6.7 Future Development

In its current form, the BuscTrad system is a fully functional tool that can be (and is being) used by Latin American clinicians to access MEDLINE-indexed content. However, there is considerable room for improvement. The user interface, while functional and useful, is quite basic compared to that of other search systems. While BuscTrad does support all of the advanced query syntax features of PubMed, there are no user interface elements to assist novice users in discovering such features. Furthermore, the current result presentation views do not allow the user to change the sort order of their results (e.g., by author, title, etc.), and only display results in descending order according to their date of publication. As such, BuscTrad's future development will include significant changes to its user interface for both the query and results display screens.

On the query screen, we plan to add controls to make it easier to limit searches by date range, publication type, and MeSh heading. Additionally, we wish to

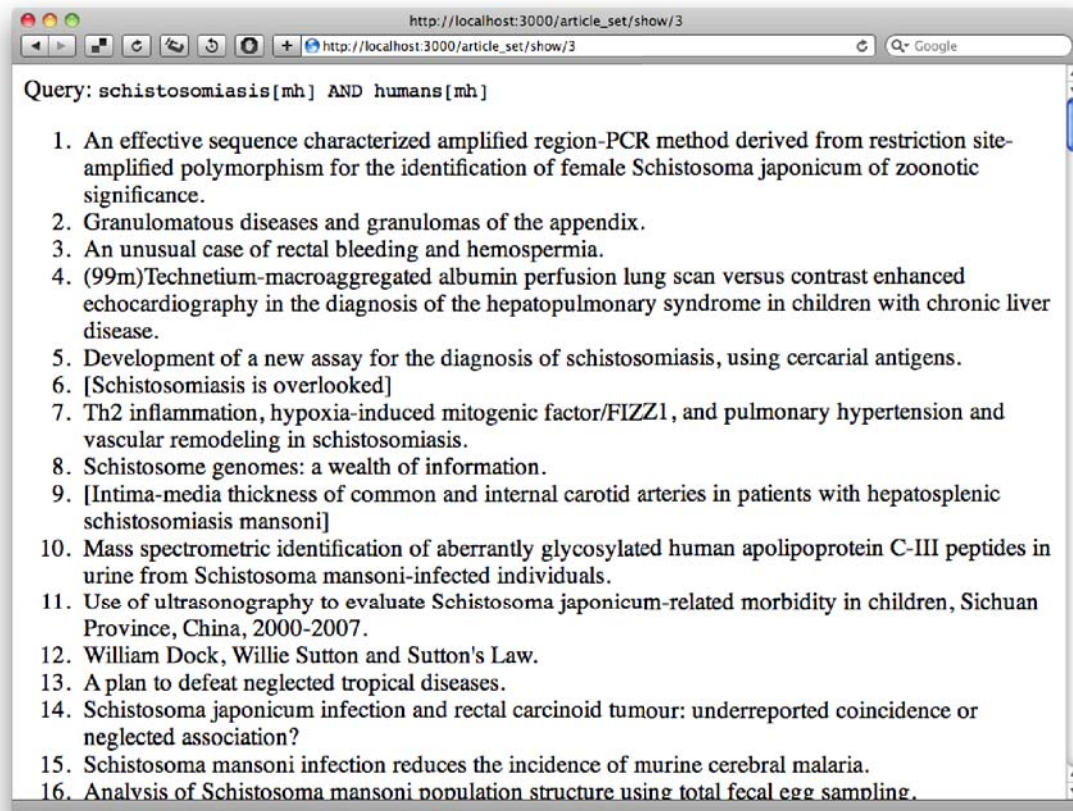


Figure 6.11: The contents of a pre-loaded set. Note that the original query remains linked to the set of articles.

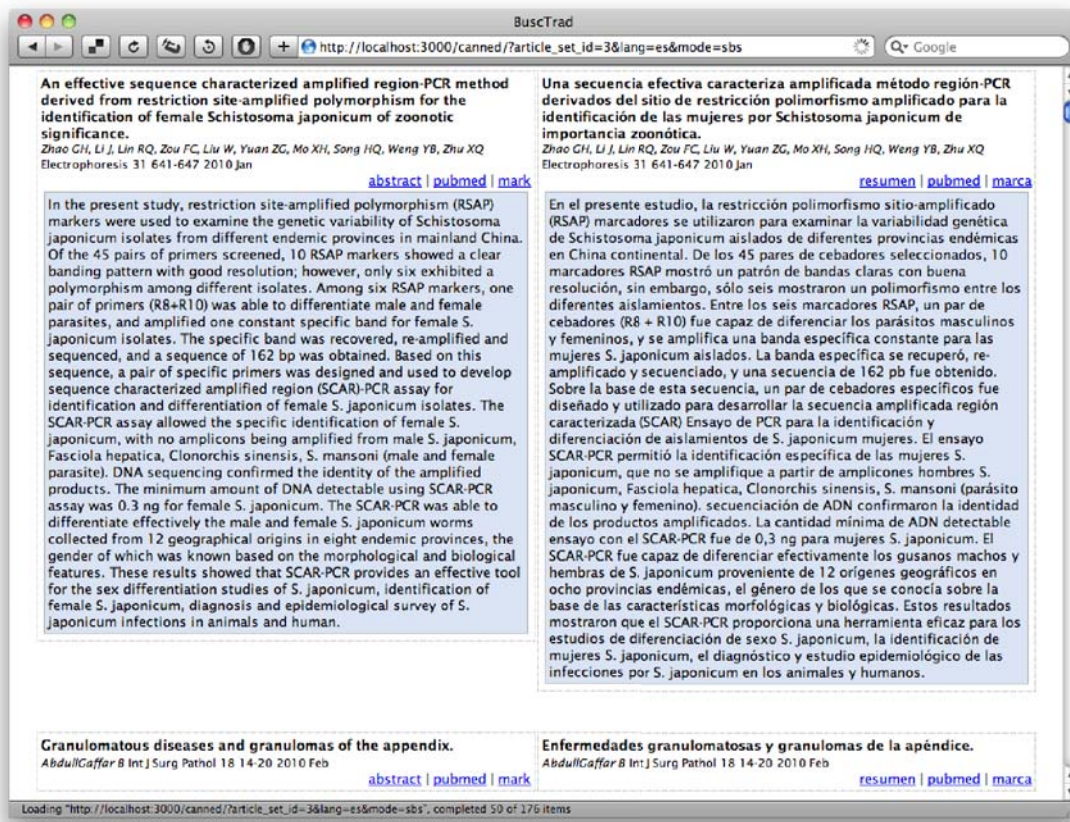


Figure 6.12: BuscTrad can display the contents of a pre-loaded article set using any of the result presentation modes described in Section 6.5.

Merge

Set One: Count:

Set Two: Count:

Title:

Figure 6.13: Investigators may create new article sets by merging two existing ones. In this screenshot, the investigator is creating a new article set that will include 30 randomly-selected articles from the set “Schistosomiasis set” and 30 randomly-selected articles from the set “Monkeys.” The new set will be named “Monkeys & Bilharzia,” and, once created, may be treated as a full article set in its own right.

add ways for users to limit their results by the availability of full-text articles. As discussed in chapter 2, one major issue facing Latin American clinicians and researchers is that of being able to access the full text of journal articles. As such, we also plan to include controls in BuscTrad's query screen to limit results to articles whose full text is freely available, either from PubMed Central or in some other open-access format; we also plan to add controls to the result display screen to highlight or otherwise filter search results falling into this category. We may also feature links or instructions to programs such as the WHO's HINARI that are designed to assist researchers and clinicians from low-income countries in accessing full-text medical articles.

Currently, BuscTrad queries must be entered in English. At some point, we wish to experiment with query translation; as discussed in chapter 4, this is an unproven technique, but one that we feel may be of value to our users. If nothing else, we may be able to construct some sort of query-building assistant that would assist users with weaker English in constructing their queries.

On the result display screen, we wish to add interface features that will allow users to control the order in which their results are displayed. Specifically, we wish to experiment with navigation facets¹¹ derived from article MeSH index terms and other relevant features (publication type, journal, etc.). As previously mentioned, we also wish to somehow make more prominent articles whose full text is freely available. This may take the form of a navigation facet, color highlighting, or something else entirely.

In addition to these user interface enhancements, we wish to expand the system to be able to accommodate languages other than English. With the exception of the

¹¹ "Faceted navigation" is a result presentation technique in which users are able to filter their search results according to a set of domain-specific metadata[71]. For example, an art search engine might allow users to filter their results according to work type (painting, sculpture, etc.), media (oil paint, watercolor, lithograph), canvas size, style (impressionist, surrealist, etc.), time period (Victorian, Georgian, etc.), geography, and so on.

result presentation interfaces, the entire system is currently language-agnostic, so it is essentially a matter of translating the various user interface features into additional languages. We currently have partial support for Arabic, including support for right-to-left text display, and are ensuring that our future development keeps in mind the need for localization and internationalization support.

Chapter 7

FREDO: FRamework for Evaluation

Design and Operation

As will be discussed in chapter 8, the evaluative component of this dissertation consisted of a user study that sought to investigate the effects of the previously-described BuscTrad system on user behavior. Our subjects were practicing Latin American clinicians, and so we therefore had to conduct all aspects of the study via the Internet. In order to conduct this study, we needed to design and build infrastructure to guide subjects through the experimental protocol, as well as to collect the actual data.

Our protocol consisted of a variety of interactions that subjects would need to undergo, including questionnaires, static instructional pages, and a relevance judgment task. The order in which the subjects were to experience these steps was to vary from subject to subject—the protocol used a Latin Square design to control for order effects between certain steps, whereas other steps were to occur in a linear order. The questionnaires and static information pages needed to be presented in both English and Spanish, and the questionnaires as written contained somewhat complex logical dependencies between questions (“if the answer to question #2 is

'yes', display question #3 and #4").

There exist numerous software packages that facilitate electronic data collection for use in research studies, ranging from relatively simplistic systems such as SurveyMonkey¹ to complete clinical trial management systems such as Vanderbilt University's REDCap [193]. While planning this study, we considered several options, and ultimately decided to build our own evaluation management system. Existing systems were either too simplistic, offered insufficient control over the visual display of study materials, had inadequate support for localized content, or could not be extended to support more complex data collection methodologies (such as that seen in our relevance judgment task).

The tool we ultimately constructed is called FREDO ("FRamework for Evaluation Design and Operation"). FREDO is designed to serve as a generalized framework for any sort of electronic data-collection process. It supports complex surveys with localized content, and can model relatively complex evaluation protocols. Most importantly, it is designed to be easy to use in conjunction with external data collection tools or to integrate into larger software environments.

7.1 Evaluation plans

The basic logical unit of FREDO is the "evaluation plan." Investigators using FREDO to conduct evaluations must first model their experimental protocols as an evaluation plan, which is composed of an ordered series of steps ("evaluation plan steps") through which subjects must be guided. Each "step" consists of one part of the study: a survey or case report form that must be filled out, an instruction sheet or consent form that must be viewed, an external site that must be visited, and so on. An evaluation plan, then, consists of a set of step definitions along with a set

¹SurveyMonkey.com, Palo Alto, CA

of instructions to about the order in which the subjects are to execute the steps.

FREDO evaluation plans may include three different types of steps:

1. Surveys (questionnaires, case report forms, etc.)
2. Static pages (landing pages, study information sheets, instructional pages, etc.)
3. External URLs (any other data-collection mechanism)

An evaluation plan may contain any number or combination of steps. By default, steps occur in a linear manner based upon the order in which they are defined; however, when developing up an evaluation plan, the investigator may choose group several steps together and randomize the order in which they are presented to subjects. In addition to block-level randomization, FREDO supports simple Latin Square designs.²

Evaluation plans are loaded into FREDO using a very simple custom programming language. While this approach may at first seem somewhat outdated and user-unfriendly, it actually has several advantages over a more “modern,” menu-driven user interface. The first important advantage is that FREDO evaluation plans, consisting as they do of a simple text file, are inherently easy to share between researchers, and to modify from study to study. They are, in essence, executable descriptions of a study protocol, and as such can serve as written records of what precise steps the subjects enrolled in a particular study underwent. This means that FREDO studies are inherently easier to reproduce than studies conducted using other toolkits. Consider two sites, *A* and *B*. *A* has recently finished conducting a user study, and *B* wishes to attempt to reproduce their results. Using FREDO, all *B* needs to do is obtain a copy of *A*’s evaluation plan script, and

² See [137] and [194] for further discussion of Latin Square designs in information retrieval evaluation.

Listing 7.1: A simple evaluation plan.

```
eval_plan('Demo Eval Plan') do |e|
  e.page('start page', :en)
  e.survey('demographics', :en)
  e.external('http://www.ohsu.edu/mystudy')
  e.page('finish page', :en)
end
```

load it into their server. They will immediately have a complete (and, most importantly, actionable) record of precisely what the subjects in site *A*'s study experienced. Without FREDO, the investigators at site *B* would have to manually re-create site *A*'s protocol as best they could from whatever records happened to be available.

The language FREDO uses to encode evaluation plans is designed to be very easy to use. It is implemented as a Ruby library, so when programming a FREDO evaluation plan, the investigator may do anything that they would ordinarily be able to do with Ruby, including making use of control structures like loops and conditional statements, as well as other common language features such as local variables and recursion. A simple evaluation plan is shown in Listing 7.1.

This evaluation plan features two static pages as well as a survey and an external resource. Before an evaluation plan may use resources such as static pages or surveys, they must first be loaded into the FREDO system (see sections 7.5, 7.4, and 7.6). External resources do not need to be pre-loaded, and are simply specified using the resource's URL. A subject participating in this evaluation, then, would first view the page entitled "start page," then fill out the "demographics" survey, then be taken to `http://www.ohsu.edu/mystudy`, and would finally end up at the page entitled "finish page." The evaluation plan further specifies that FREDO display the English-language version of the pages and survey.

A more complex evaluation plan is shown in Listing 7.2. A subject participating in this study would visit three external sites, and the order in which they would

Listing 7.2: A more complex evaluation plan. Note the use of randomized steps.

```
eval_plan('Complex Eval Plan') do |e|  
  e.page('start page', :en)  
  e.survey('demographics', :en)  
  e.randomize do |r|  
    r.external('http://mystudy.com/taskA')  
    r.external('http://mystudy.com/taskB')  
    r.external('http://mystudy.com/taskC')  
  end  
  e.survey('followup', :en)  
  e.page('finish page', :en)  
end
```

do so would be randomized. All other aspects of the study (the demographic and follow-up surveys, etc.) would occur in a linear manner.

There is no limit to the number of steps an evaluation plan may contain. There is also no limit to the number of randomization or Latin Square blocks that a plan may contain. However, FREDO evaluation plans may not contain nested randomization or Latin Square blocks. Note that the system logs the order in which each subject participating in an evaluation plan encounters each step. So, for example, if an evaluation plan calls for steps 4, 5, and 6, to be randomized, the system will record—on a per-subject basis—whether the subject’s exposure to the steps was in the order $5 \rightarrow 4 \rightarrow 6$, or whether it was $6 \rightarrow 4 \rightarrow 5$, etc. This enables investigators to ensure that their randomizations or Latin Squares are functioning appropriately.

Although the evaluation plan syntax may initially appear somewhat complex to non-programmers, it is in practice simpler than it looks. It is designed to be as simple and minimalist as possible, with the the goal that study planners be able to make use of it with minimal support from informaticians or system administrators.

7.2 Collectors

Since by definition FREDO studies are conducted via the Internet, investigators must provide their subjects with the address of a “landing page” (i.e., the starting point for the study). FREDO calls these “collectors,” and one must be configured for each evaluation plan. Figure 7.1 illustrates FREDO’s user interface for configuring collectors.

FREDO allows investigators to configure multiple landing URLs for a single evaluation plan. This feature’s primary purpose is to allow investigators to set up multiple avenues through which subjects may participate in their studies, and to easily distinguish thereafter the origin of their subjects.

Consider the case of an investigator planning on enrolling subjects from multiple sites into her study. One option would be to set up separate evaluation plans—one for each study site—as described in the previous section. However, this has the disadvantage that any changes to the protocol will need to be precisely duplicated across each site’s evaluation plan, which introduces the possibility for subtle and difficult-to-diagnose errors.

Instead, the investigator could simply configure a collector (linked to the same evaluation plan) for each study site. FREDO will then label each subject participating in a study with the collector from which they arrived at the site, thereby enabling the investigator to easily tell “where” a given subject came from. The investigator must simply provide subjects at each study site with the appropriate collector URL, and the system will take care of the rest.

FREDO provides an administration interface that allows investigators to easily see how many subjects have participated in their evaluation from that collector, as well as retrieve collected data from that controller’s subjects from any of the plan’s steps (see Figure 7.2). FREDO also logs each subjects’ progress through the evaluation plan (Figures 7.3 and 7.4), and this data is available to administrators

from the same screen as the higher-level log data.

7.3 System Flow

FREDO's primary role is to keep track of a study's subjects as they progress through the various steps of a study, and to ensure that each step occurs in the correct sequence. The process begins when a subject initially arrives by following a link to a collector. The subject could have received the collector link via email, copied it off of a flier or information sheet, or obtained it in any number of other ways. Whatever their origin, the first thing the system does with new visitors to a collector URL is assign them a unique subject identification number, which will follow them throughout the course of their participation in the study.

After assigning the new arrival a subject identification number, FREDO next must determine which evaluation plan the subject is seeking to participate in. This is a simple matter of linking the collector to its evaluation plan. Once FREDO has identified the evaluation plan, it then calculates the order in which the plan's steps are to be presented to the subject. As described earlier, this may involve randomizing (or "shuffling") the order of some steps, or consulting a Latin Square. FREDO uses the standard Fisher-Yates/Knuth algorithm to perform step randomization[195].

Once FREDO has calculated a subject's step order, it is then a simple matter of directing the subject through each step. Upon each request, the system looks up the current subject's step order, and sends them to the next uncompleted step in their plan. Since all steps in a FREDO evaluation plan are accessible via a URL, FREDO uses standard HTTP redirection (described in Section 6.1) to redirects the subject's web browser to the appropriate URL as specified by the step. Once this redirection occurs, the subject is temporarily out of FREDO's control. It is the responsibility of the software behind the step itself— the specific survey, external URL, etc.—

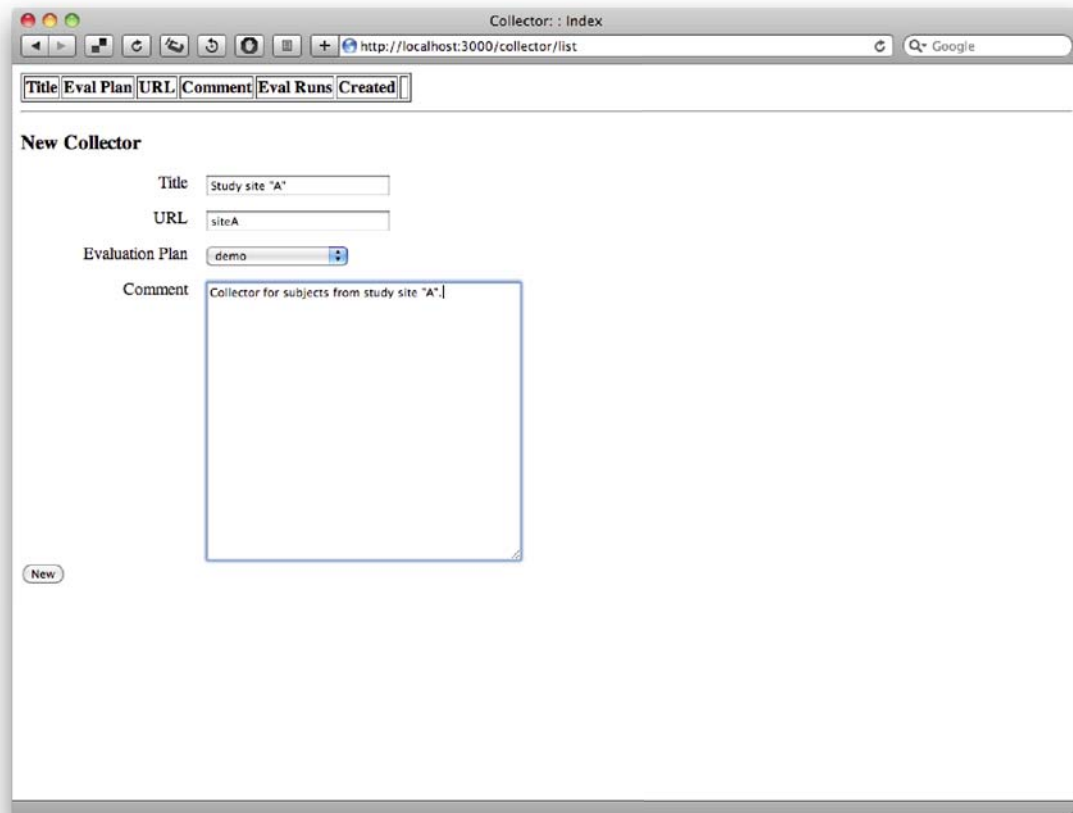


Figure 7.1: New collectors are set up via a web interface. Investigators have complete control over the URL pattern that they wish to be assigned to their evaluation plan. In this case, subjects wishing to participate in the protocol described in the “demo” evaluation plan will be able to do so by visiting `http://server address/c/siteA`.

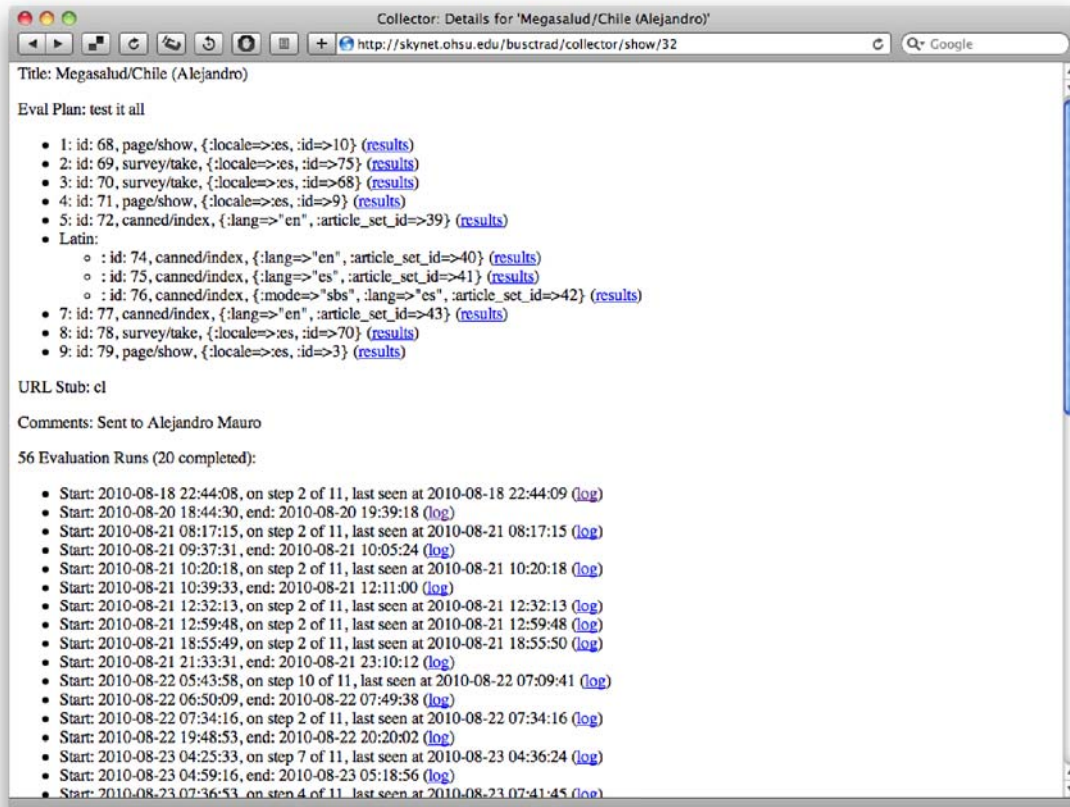


Figure 7.2: Study administrators may access data about how many subjects have participated in a given evaluation plan from a particular collector, and from the same screen may easily access any data that has been collected to date.

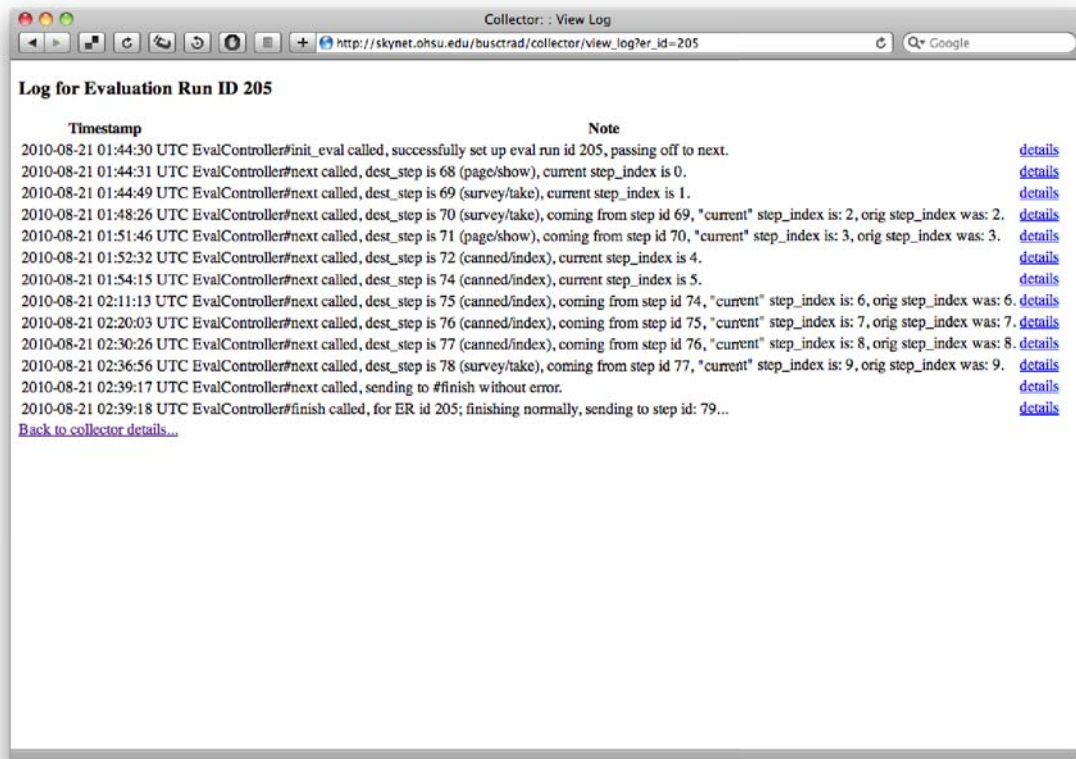


Figure 7.3: FREDO logs data about each subject's participation, including when they transitioned from one step within an evaluation plan to another.

Collector: : Log Details

Created At: 2010-08-21 01:44:30 UTC
 ER ID: 205
 Note: EvalController#init_eval called, successfully set up eval run id 205, passing off to next.
 Params Hash:

Header	Value
SERVER_NAME	skynet.ohsu.edu
rack.url_scheme	http
rack.run_once	false
rack.input	#
HTTP_ACCEPT_ENCODING	gzip, deflate, sdch
HTTP_USER_AGENT	Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US) AppleWebKit/533.4 (KHTML, like Gecko) Chrome/5.0.375.126 Safari/533.4
PATH_INFO	/evaluation/init_eval
rack.errors	#
HTTP_ACCEPT_LANGUAGE	es-ES,es;q=0.8,en-US;q=0.6
HTTP_HOST	skynet.ohsu.edu
SCRIPT_NAME	/busctrad
SERVER_ADDR	137.53.250.53
SERVER_PROTOCOL	HTTP/1.1
REMOTE_ADDR	190.101.45.32
SERVER_SOFTWARE	Apache/2.2.8 (Ubuntu) Phusion_Passenger/2.2.15 PHP/5.2.4-2ubuntu5.10 with Suhosin-Patch proxy_html/3.0.0 mod_perl/2.0.3 Perl/v5.8.8
rack.request.query_hash	eval_plan_id25collector_id32
rack.multithread	false
rack.version	10
HTTP_COOKIE	utmz=67595343.1280803002.2.2.utmcsr=google utmccn=(organic) utmcmd=organic utmctr=ohsu; WT_FPC=id=10.76.6.140-435913776.30086869:lv=1280788610362:ss=1280788602088; utma=67595343.159082912.1277738286.1277738286.1280803002.2
HTTP_ACCEPT_CHARSET	ISO-8859-1,utf-8;q=0.7,*;q=0.3
rack.multiprocess	true
REQUEST_URI	/busctrad/evaluation/init_eval?collector_id=32&eval_plan_id=25
DOCUMENT_ROOT	/var/www/
SERVER_PORT	80

Figure 7.4: For each subject interaction, FREDO logs detailed technical information including data about their web browser configuration.

to do with the subject whatever it is meant to do (collect information, provide instructions, etc.), and, when it is finished, return the subject to FREDO.

When FREDO redirects a subject to a new step, the subject does not arrive “out of the blue:” FREDO sends with the subject several vital pieces of data, including the aforementioned subject identification number, the identification number for the collector from which the subject initially arrived, and the step’s internal FREDO identifier. These data are sent in the form of “url parameters”: key-value pairs that are appended to the URL sent as part of FREDO’s HTTP redirect. These data have two purposes. The first is to provide the step itself with context about the subject and their position within the evaluation plan. By including the subject identifier, the step can associate any data it collects with the subject who generated the data, thereby enabling the investigator to easily link the data from all of the evaluation plan’s steps together.

The second purpose behind including contextual information is that it provides a way for FREDO to identify subjects returning from externally-hosted steps. When a step is ready to send the subject back to FREDO to continue on with the evaluation (e.g., the subject has completed a survey, or is otherwise ready to go on to the next data collection task), all the step has to do is redirect the subject’s browser back to a specific FREDO url (referred to as the “next” url³), and as long as they include the subject identification number, FREDO will be able to determine where next to send the subject. If the subject has completed all the steps, FREDO will close out their session and redirect them to to the plan’s final step. Figure 7.5 illustrates this process.

³ By default, `http://server address/next`

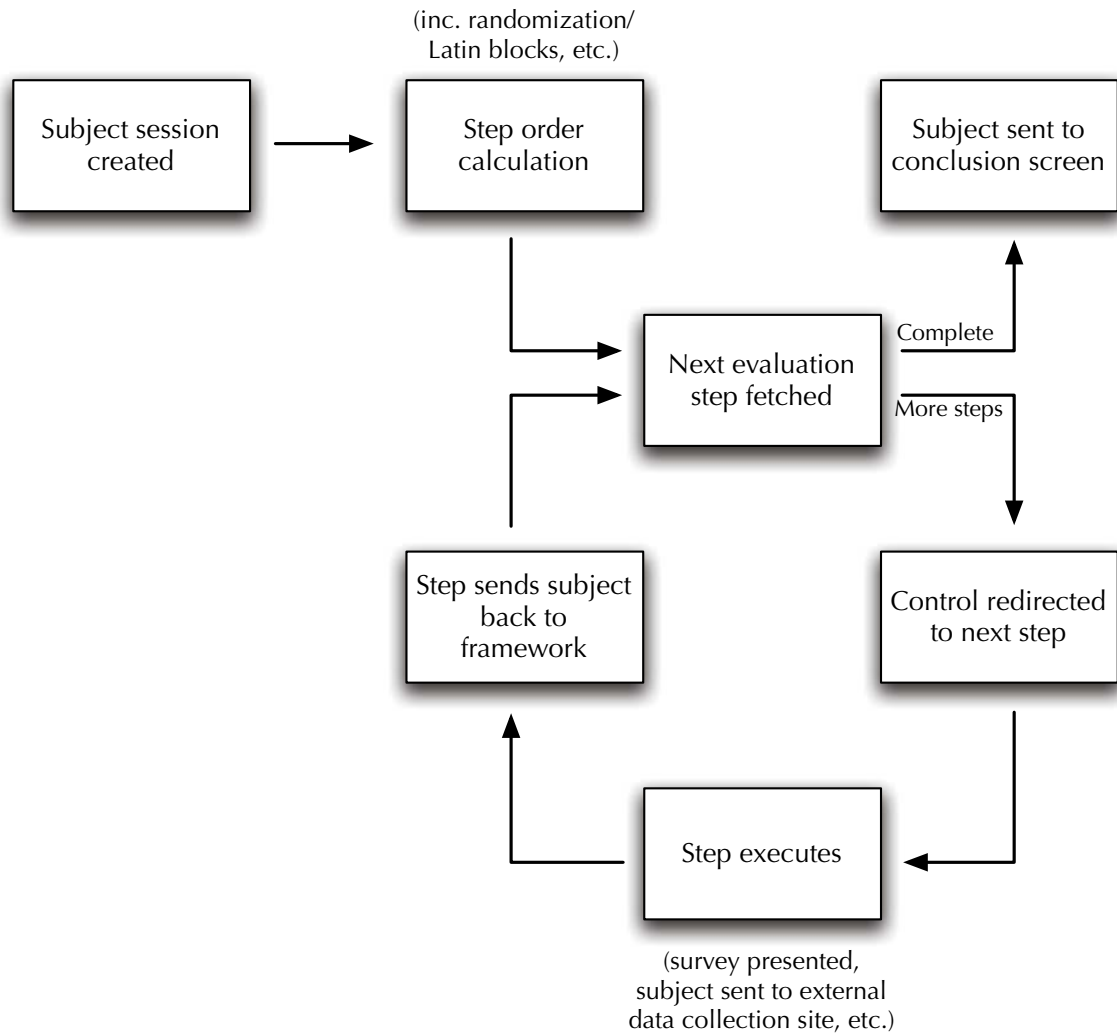


Figure 7.5: FREDO guides subjects through evaluation plans one step at a time. After each step, FREDO checks to see whether there are more to complete; if so, it hands control over to the next step, which eventually hands control back to FREDO, and the cycle repeats. When there are no more steps for the subject to complete, the subject's record in the database is marked as "finished," and the subject is sent to the final step in the plan (typically a static page, or an external site).

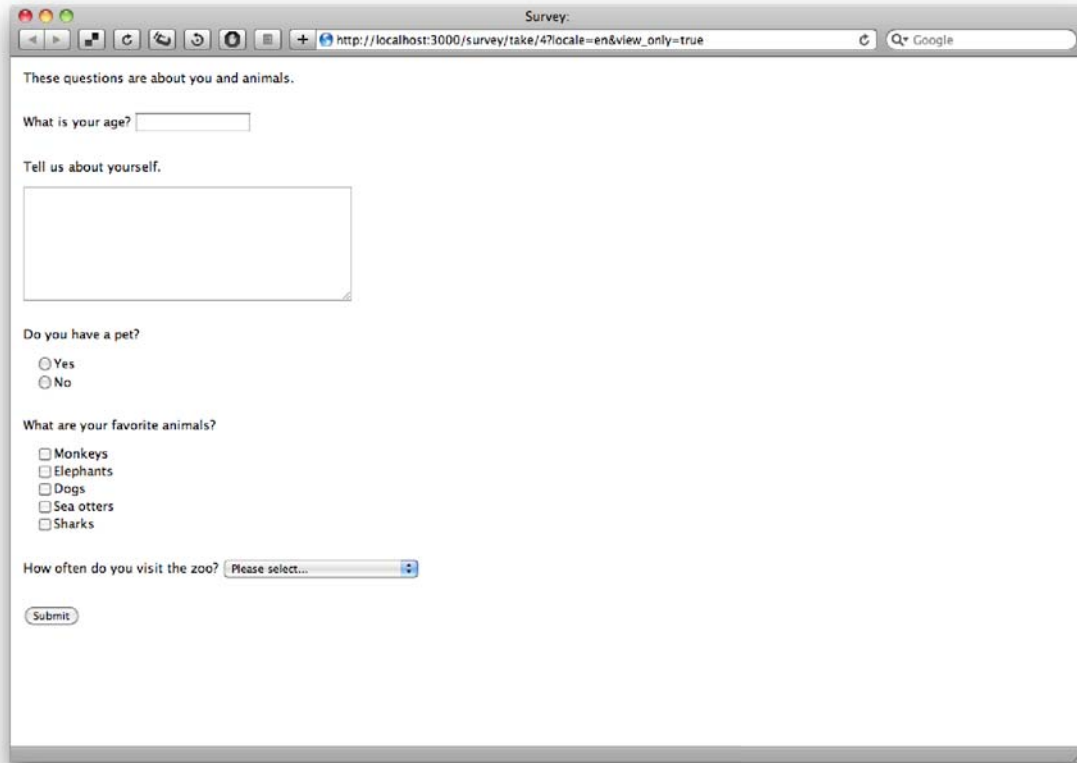
7.4 Surveys

One of the primary use cases for the FREDO system is the management of studies that involve collecting data from human subjects. Often, this is done by means of surveys (sometimes referred to as questionnaires, or case report forms, etc.). While one could easily use a third-party survey system such as SurveyMonkey in conjunction with FREDO, investigators wishing for a more integrated experience— or for more advanced features and localization capabilities— would be better served by FREDO’s built-in survey system.

A FREDO survey is an ordered set of *questions*, each of which have textual *labels*. Questions in a survey may have labels in any number of languages (locales). Questions may be of six different types:

1. Short text fields
2. Long text fields (“text areas”)
3. Radio buttons
4. Check boxes
5. Drop-down selection menus
6. Text banners

Figure 7.6 shows an example of a simple FREDO survey that includes questions of several types. Text fields capture textual responses from subjects, and may be either “short” or “long” depending on how much of a response is expected. Radio buttons allow for “single-selection” multiple-choice questions— subjects must pick one and only one of the answers. Check boxes, on the other hand, allow subjects to select multiple answers (“check all that apply...”). Drop-down menus



Survey:

http://localhost:3000/survey/take/47?locale=en&view_only=true

These questions are about you and animals.

What is your age?

Tell us about yourself.

Do you have a pet?

Yes
 No

What are your favorite animals?

Monkeys
 Elephants
 Dogs
 Sea otters
 Sharks

How often do you visit the zoo?

Figure 7.6: FREDO surveys may feature several different type of question, including multiple choice, free-text, and so forth.

provide an alternate mechanism for single-choice answers to multiple-choice questions, and have the added advantage that they may be presented in the middle of their labels (see Figure 7.7).

“Text banners” are questions without answers, and as such are intended as a way for survey designers to insert explanatory text or instructions into their survey between and around the regular questions. For example, in the screenshot shown in Figure 7.6, the text at the very beginning of the survey (“These questions are about...”) is a text banner.

The FREDO system stores survey specifications, questions, and subjects’ answers to survey questions in a relational database. While investigators and survey authors could, in theory, directly access their surveys’ results directly using SQL, the system provides a graphical interface for reviewing and exporting survey re-

I thought that the system was the easiest one to use.

The system was the most efficient one to use.

I was fastest at using the system.

I liked using the system the most.

Figure 7.7: One of the advantages of drop-down menus is that they may be interpolated within their labels.

sults. This system is discussed in Section 7.4.3.

7.4.1 Survey Syntax

As with evaluation plans, investigators load their survey instruments into FREDO using a custom programming language. Many modern survey packages (SurveyMonkey, etc.) offer graphical form building tools; FREDO lacks such an interface. While these graphical tools are often excellent for simple forms, they can be quite cumbersome for complex surveys. Consider the case of an investigator planning a survey that contains many similar questions, or questions with common sets of possible answers. Many graphical form builders require a great deal of repetitive (and error-prone) clicking to set up such a survey. Using FREDO's scripting approach, however, our investigator may simply use her text editor's "copy" and "paste" commands. Depending on the circumstance, however, she may be even better off using taking advantage of the fact that FREDO's survey-building syntax is built on top of a full programming language. As such, she may use any number of different programming techniques, such as loops or other such control structures; alternatively, she may define a custom function for repetitive tasks and re-use it as a "macro."

Another important advantage to a scripting-based approach to survey construction is that it makes changes easier to implement. Imagine that our inves-

tigator, upon reviewing an initial draft of the survey instrument, decided to make a wording change that would affect many of the survey's questions. Using most graphical form builders, this would involve yet more repetitive clicking through numerous dialog boxes and screens. Using FREDO's approach, however, our investigator could simply use her text editor's "find and replace" functionality.

A third advantage of the scripting-based approach is that it makes collaboration easier. Collaborators may simply share a single text file containing the script for the survey among themselves, revising as necessary, much as they would any other document. Since the source file for the survey is itself a simple text file, there exist many different software packages that can help our investigator share the survey source file among a group of collaborators, and to track any changes that may be made.

The syntax for describing a survey is somewhat similar to that for describing evaluation plans. Listing 7.3 on page 117 contains the syntax describing the survey shown in Figure 7.6, and illustrates several different question types. Note that each question contains label and answer text in both English and Spanish; if at a later point in time the investigator decides to add a third language to the survey, it is a simple matter of adding a few lines to the survey specification.

This sample syntax also illustrates three ways that survey authors may specify the answers to their multiple-choice questions. The "pets" question ("Do you have any pets?", line 16 in Listing 7.3) has two possible answers: "Yes" and "No." The author of this survey has chosen to code "Yes" answers with the number one, and "No" answers with the number two. This is entirely arbitrary, and survey authors may choose whatever coding scheme they wish.

The "pets" question is relatively simple, so manually specifying each possible answer (along with its coding pattern) in this manner is feasible. However, in the case of questions with more possible answers, it would be quite tedious to specify

every possible answer (along with its labels, in however many languages as may be necessary) manually. As such, the FREDO survey system has several built-in macros to simplify this task.

For an example of this, consider line 33 in Listing 7.3, which illustrates a second way to add multiple answers to a question. This question is making use of the `answer_array` macro, which allows survey authors to simply provide a list of answers, which are represented as sets of key-value pairs wherein the keys are the locale codes and the values are the locale-specific texts of the answers. This macro makes entering lists of answers much easier and faster, but it does so at a cost. Using the manual approach, the survey author has complete control over how the various answers are coded. When using the `answer_array` macro, however, the system will automatically encode each answer with an ascending integer. So, the first answer will be coded as “one,” the second as “two,” and so on.

A third way to add answers to questions is shown on line 47 of Listing 7.3. The FREDO survey system features macros that add English and Spanish versions of several very common “answer sets.” In this case, the author is using the `frequency_scale` macro, which adds several possible answers involving frequency (“Every day or almost every day,” “Once or twice per week,” etc.). The macro also supports author-configurable text for a “Never” option, as shown in Listing 7.4. In this example, the question would appear to the subject as a drop-down menu containing several frequency options as well as the text “I never write English-language text at work.”

In using the `frequency_scale` macro, as with `answer_array`, the survey author gives up control over how the various answers are to be coded. There are several other standard macros. One example is the `yesno` macro, which is a very succinct way to add a “yes or no” response to a question (as shown in the several lines of code beginning on line 20 in Listing 7.3). Another important

macro is the `likert` macro, used to automatically add Likert-scale answers. An author using the `likert` macro must supply text labels to be placed at the “low” and “high” anchors, and the FREDO system will automatically set up a range of numerical answers (see Listing 7.5 and Figure 7.8). In this example, the author is using the system’s default of a five-point Likert scale. Authors may specify a different number of steps, and may also specify the interval between steps.

Listings 7.3, 7.5, and Figure 7.8 also demonstrate the use of the `instructions` macro, which provides authors a more natural way to add text banners containing instructional or advisory text to their surveys.

Listing 7.3: A simple survey, illustrating several question types.

```
1 survey('Zoo Survey') do |s|
2   s.instructions('instruct', {
3     :en => 'These questions are about you and animals.',
4     :es => 'Estas preguntas son acerca de usted y los animales.'
5   })
6   s.question('age') do |q|
7     q.type SimpleTextField
8     q.label :en, 'What is your age?'
9     q.label :es, '¿Cuántos años tiene usted?'
10  end
11  s.question('personal') do |q|
12    q.type TextArea
13    q.label :en, 'Tell us about yourself.'
14    q.label :es, 'Díganos acerca de usted.'
15  end
16  s.question('pets') do |q|
17    q.type RadioButton
18    q.label :en, 'Do you have a pet?'
19    q.label :es, '¿Tiene una mascota?'
20  q.answer('1') do |a|
```

```
21     a.label :en, 'Yes'
22     a.label :es, 'Si'
23     end
24     q.answer('2') do |a|
25         a.label :en, 'No'
26         a.label :es, 'No'
27     end
28 end
29 s.question('animal') do |q|
30     q.type CheckBox
31     q.label :en, 'What are your favorite animals?'
32     q.label :es, '¿Cuáles son sus animales favoritos?'
33     q.answer_array(
34         [
35             {:en => 'Monkeys', :es => 'Monos'},
36             {:en => 'Elephants', :es => 'Elefantes'},
37             {:en => 'Dogs', :es => 'Perros'},
38             {:en => 'Sea otters', :es => 'Chungungos'},
39             {:en => 'Sharks', :es => 'Tiburónes'}
40         ]
41     )
42 end
43 s.question('zoo_freq') do |q|
44     q.type DropDownList
45     q.label :en, 'How often do you visit the zoo?'
46     q.label :es, '¿Con qué frecuencia visita el zoo?'
47     q.frequency_scale
48 end
49 end
```

Listing 7.4: A simple survey, illustrating several question types.

```

s.question('work_eng_write_freq') do |q|
  q.type DropDownList
  q.label :en, 'At work, how often do you <u>write</u> English-language
    text?'
  q.label :es, 'A su trabajo, ¿con qué frecuencia escribe textos en
    idioma Inglés?'

  q.frequency_scale(true, {
    :en => 'I never write English-language text at work.' ,
    :es => 'No tengo que escribir texto en idioma Inglés.'
  })

```

Listing 7.5: An example of a simple use of the likert and instructions macros.

```

s.instructions('non_work_self_english_instructions', {:en => 'When you
  encounter the English language in <u>non-professional</u> contexts,
  how would you rate your level of proficiency at:', :es => "<strong>
  Respecto al uso del idioma Inglés <u>en contextos no émdicos</u>,
  ¿como calificaría usted su nivel de dominio de:</strong>"}))

s.question('non_med_self_eng_read') do |q|
  q.type DropDownList
  q.label :en, '<u>reading</u> English?'
  q.label :es, "<u>lectura</u> en Inglés?"

  low_anchor = {:en => 'Cannot read any English', :es => 'No sabe leer
    Inglés'}
  high_anchor = {:en => 'Perfect English reading ability', :es => '
    Capacidad de lectura en inglés perfecta'}

  q.likert(low_anchor, high_anchor)

```

end

When you encounter the English language in non-professional contexts, how would you rate your level of proficiency at:

reading English? ✓ Please select...

writing English?

speaking in Eng

1 - Cannot read any English

2

3

4

5 - Perfect English reading ability

Figure 7.8: A drop-down question using answers generated via the likert macro, as shown in Listing 7.5.

7.4.2 Question Dependencies

The FREDO survey system also supports dependencies between questions. In other words, it is possible to configure a survey wherein the answer to one question affects whether or not other questions are visible to subjects. The syntax for specifying question dependencies is simple, and is shown in Listing 7.6. The author of the example shown has specified two radio button questions, and has used the `yesno` macro to automatically populate each question with a pair of answers: “Yes” (coded as “1”) and “No” (coded as “2”).

By placing the second question’s declaration inside a `dependent` block (see line 7 in Listing 7.6), the author specifies that the second question should only appear if the subject’s answer to the first question is “1” (i.e., “Yes”). If the subject changes their answer to the first question (e.g., to “No”), the second question will disappear from the subject’s view of the survey.

The mechanism for determining whether or not to show a dependent question is currently quite simplistic, and can only handle basic equality predicates. In other words, authors may only specify a single value to trigger the dependency (as opposed to a range of values, or a boolean inequality). In practical terms, this means that certain types of logic cannot currently be encoded. For example, in the example in Listing 7.6, the author asks the question: “Are you over 18?” If the answer is “Yes,” the dependency is triggered. Consider instead the case where the survey author had wished to ask the subject’s age, and display the follow-up question in the event that the age was *greater than* 18. FREDO’s survey syntax does not currently support this sort of complex logic. As currently implemented, it would be non-trivial to add support to FREDO’s survey syntax for more complex logic, although it would by no means be an insurmountable obstacle.

One common use-case for dependent questions is the “Other - please specify” pattern (see Figure 7.9), in which a list of possible answers ends with an “Other”

Listing 7.6: An example of a question dependency. The second question will only appear if the answer to the first question is “1,” which is how the `yesno` macro encodes “Yes.”

```

1
2 survey('Test Dependency') do |s|
3   s.question('age') do |first_q|
4     first_q.type SimpleTextField
5     first_q.label :en, 'Are you over 18?'
6     first_q.yesno
7     s.dependent(first_q, '1') do
8       s.question('vote') do |second_q|
9         second_q.type RadioButton
10        second_q.label :en, 'Are you registered to vote?'
11        second_q.yesno
12      end
13    end
14  end
15 end

```

option that, when selected, causes a text field to appear in which the subject may enter a value. Because this is such a common question format, the FREDO survey system contains an `other` macro (see line 34 in Listing 7.7) that authors may use to easily add an “Other” option with a corresponding text entry box to their multiple choice questions. Listing 34 illustrates `other`’s use in conjunction with a manually-specified answer set; the `answer_array` macro can also be instructed to automatically add an “Other” option.

7.4.3 Administration

One important aspect of survey-based data collection is the retrieval from the system of the subjects’ responses. Each survey in a FREDO installation features an administrative screen, from which investigators may view their subjects’ responses (see Figure 7.11), as well as download those responses in a standard comma-separated format (CSV) that may be easily imported into any statistical analysis package. From this screen, they may also “preview” their surveys, to ensure that the questions and logic are correct, and modify existing questions and answers.

Listing 7.7: An example of chained dependencies implemented using the `other` macro.

```
1
2 s.question('use_social_networking') do |q|
3   q.type RadioButton
4   q.label :en, "Do you use social networking sites (Facebook, Orkut,
5     MySpace, Twitter, etc.)?"
6   q.label :es, '¿Utiliza un sitio de redes sociales como Facebook, Orkut
7     , MySpace, Twitter, etc.?'
8   q.yesno
9
10  s.dependent(q, '1') do
11
12    s.question('pref_social_networking') do |q2|
13      q2.type DropDownList
14      q2.label :en, 'What network do you use most often?'
15      q2.label :es, '¿Cual es la red social que utiliza con más
16        frecuencia?'
17
18      q2.answer('1') do |a|
19        a.label :en, 'Facebook'
20      end
21      q2.answer('2') do |a|
22        a.label :en, 'MySpace'
23      end
24      q2.answer('3') do |a|
25        a.label :en, 'Orkut'
26      end
27      q2.answer('4') do |a|
28        a.label :en, 'LiveJournal'
29      end
30      q2.answer(5) do |a|
31        a.label :en, 'Twitter'
32      end
33      q2.answer('6') do |a|
34        a.label :en, "Other"
35        a.label :es, "Otro"
36      end
37      s.other(q2, '6', 'pref_social_networking_other')
38    end
39  end
40 end
```

Do you use social networking sites (Facebook, Orkut, MySpace, Twitter, etc.)?

Yes
 No

What network do you use most often?

Other

Figure 7.9: A radio button with a dependent drop-down menu, which in turn has an “Other” option, which in turn triggers a text box question allowing the user to enter a free-text response. The drop-down menu uses the `other` macro, as shown in Listing 7.7.

Since some investigators’ collaborators may not feel comfortable using (or, for one reason or another, may not be able to use) the administration interface, users may also download the contents of their surveys as files that may be opened in Microsoft Word or other word processing systems. Finally, users of the administration interface can download a dynamically-generated “data dictionary” (see Figure 7.10), which they may use to “decode” their subjects’ responses to multiple-choice questions.

Figure 7.10 illustrates part of the administration screen for the survey shown in Figure 7.6, with the automatically-generated data dictionary visible. Note that the structured questions are all represented (while the free-text questions are not, as they do not result in coded data), and that their coding matches that specified in their survey description file (Listing 7.3). Figure 7.11 illustrates the user interface for reviewing survey results. In this figure, both the standard tabular view as well as the CSV view are enabled.

Note the drop-down menus above the table of results. These allow the user to “filter” the results display to include only results from subjects who participated in the survey from a particular collector or as part of a particular evaluation plan.

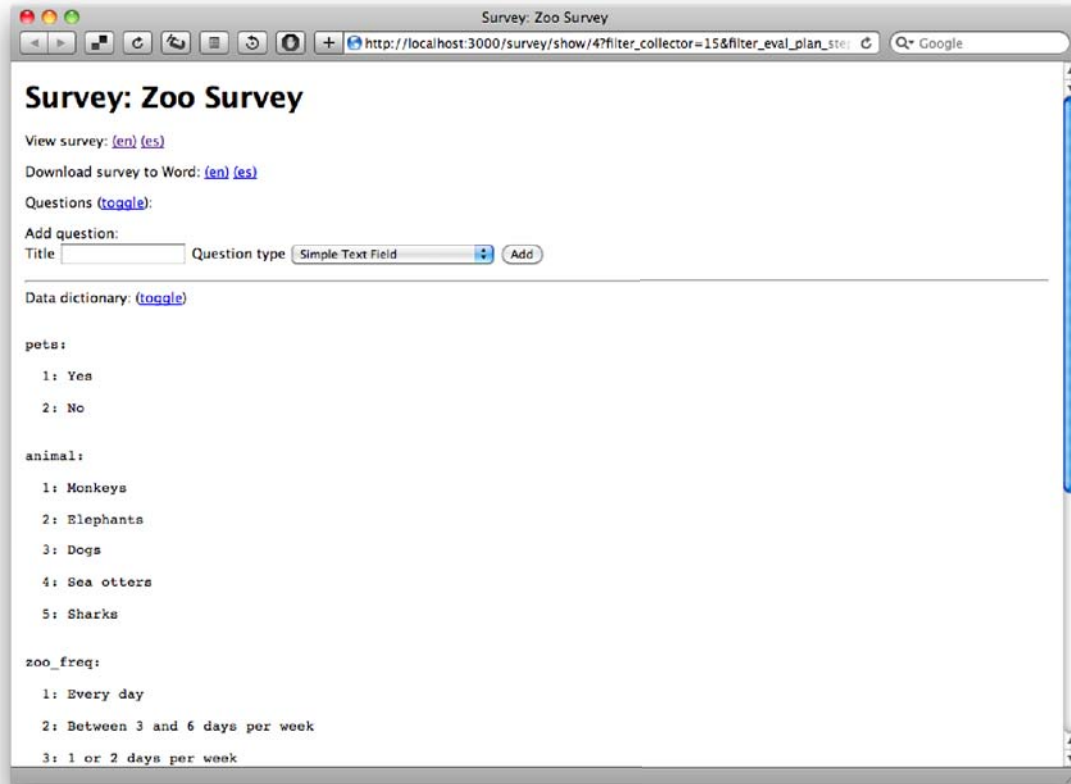


Figure 7.10: The administration screen for the survey described in Listing 7.3 and shown in Figure 7.6, with the data dictionary visible. Note that the coding scheme shown in this figure matches that specified in the survey’s description file. Clicking the “toggle” links next to each section “toggles” that section’s visibility.

This becomes important when the same survey instrument is re-used in multiple studies. Users may also enable a filter that only shows results from subjects who *finished* their evaluation plan— in other words, in the filtered mode, the table will exclude subjects who left the study part-way through their protocol. These filter settings apply to both the CSV and tabular result views.

7.5 Static Pages

In addition to questionnaires, electronic studies often need to present static information to subjects. This may take many forms: an instruction screen, an information

Survey: Zoo Survey

1: Monkeys
2: Elephants
3: Dogs
4: Sea otters
5: Sharks

zoo_freq:
1: Every day
2: Between 3 and 6 days per week
3: 1 or 2 days per week
4: One or two days per month
5: Never

Responses: [toggle](#)
demo collector 22 Only include finished eval runs? filter

CSV Responses: [toggle](#)
collector_id,evaluation_run_id,eval_plan_step_id,recorded_at,13_age,14_personal,15_pets,16_animal,17_zoo_freq
15,19,22,2010-12-28 18:43:52 UTC,29,I like animals.,2,"1, 2, 4",1
15,20,22,2010-12-28 18:44:20 UTC,97,"We didn't have "animals" back in my day...",1,"4, 5",3

[Back to List](#)

collector_id	evaluation_run_id	eval_plan_step_id	recorded_at	13_age	14_personal	15_pets	16_animal	17_zoo_freq
15	19	22	2010-12-28 18:43:52 UTC	29	I like animals.	2	1, 2, 4	1
15	20	22	2010-12-28 18:44:20 UTC	97	We didn't have "animals" back in my day...	1	4, 5	3

Figure 7.11: The administration screen for the survey described in Listing 7.3 and shown in Figure 7.6, with the responses visible in both comma-delimited (CSV) and standard tabular presentations. The CSV mode is designed to be easy to load into standard analytical packages such as Excel, SPSS, etc. Note the filter controls, which allow users of this screen to control which subjects' responses are visible.

sheet including consent data, a “wrap-up” screen to be displayed at the end of a study, and so on. While it would certainly be possible to achieve this functionality using the “text banner” survey question type (see page 7.4 in Section 7.4), this approach is somewhat limited in its capabilities. Text banner questions may only contain simple formatting instructions, and are not flexible enough for many applications.

To address this issue, FREDO contains what is in essence a simple content-management system, allowing investigators to load semi-static pages that may then be used as steps in an evaluation plan. The interface for this feature is a very simple web page, wherein page authors may enter the raw HTML (Hyper-Text Markup Language) describing the page they wish to display. Authors may load locale-specific versions of each page (e.g., an English as well as a Spanish version), and they may also load page-specific formatting instructions in the form of Cascading Style Sheets (CSS), a standard mechanism for encoding layout and style information for web pages.

In addition to simple static HTML and CSS, authors may add dynamic elements to their pages using a standard template language called eRuby⁴ (Embedded Ruby), which allows authors to insert executable code in the Ruby programming language inside an otherwise static HTML document. The server’s eRuby interpreter will, for each request, “compile” the template, thereby executing this embedded code before generating the HTML, which the server will deliver to the viewer as normal. In practice, authors of dynamic pages using this technique are using the embedded code to affect certain parts of the final page’s appearance. This technique is extremely commonly used by web programmers, and would be familiar to anyone experienced in such matters (i.e., whoever was building the web

⁴See <http://www.ruby-doc.org/stdlib/libdoc/erb/rdoc/> or <http://www.modruby.net/en/index.rbx/eruby/whatis.html> for more information on eRuby.

Listing 7.8: A simple example of using eRuby to embed dynamic elements into a FREDO page. Listing 7.9 shows the final result of this FREDO page.

```
1
2 <h1>Thank you for participating.</h1>
3
4 Two plus two is: <%= 2 + 2 %>.
5
6 <p/>
7
8 You began this study on <%= Time.now.strftime("%A, %B %d") %>.
```

content for the investigator of a FREDO-hosted study). Listing 7.8 illustrates this process.

In an eRuby document, text inside the “<%” and “%>” delimiters is treated by the server as executable Ruby code, and adding an “equals sign” (“=”) to such a code block will cause the result of the contained computation to be inserted into the document. Line 4 in Listing 7.8 illustrates a trivial example of this process. The code inside the delimiters would be executed, resulting in the number “4,” which would then be inserted (because of the equals sign) into the output. The final line, as seen by the viewer, would therefore read “Two plus two is: 4.”

Line 8 illustrates a somewhat more useful application of this technique. In this line, the author has used Ruby’s `strftime` function to perform a date formatting operation on the current date, and then causes the formatted date to be inserted into the final output. One crucial thing to realize is that the date inserted would be the date at the time of viewing, not at the time the page was initially authored. So, for example, if a subject viewed this page on November 3rd, 2010, the final output would read “You completed the study on Wednesday, November 3.” If, instead, the subject viewed the page on the 4th, the final line would read “Thursday, November 4,” and so on. Listing 7.9 illustrates this result.

This sort of customization was the intended use case behind allowing dynamic

Listing 7.9: The final output of the static page defined in Listing 7.8.

```
1  
2 <h1>Thank you for participating.</h1>  
3  
4 Two plus two is: 4.  
5  
6 <p/>  
7  
8 You began this study on Wednesday, December 29.
```

content in FREDO pages. However, there are several other potential uses. Authors of FREDO pages are able to access certain data about the context in which their pages are being viewed. For example, authors may embed into the page information about the collector from which the current subject arrived at the page (see Section 7.2), data about how far along the subject is in their evaluation plan, and so on.

One important use for this functionality is in the creation of navigation links between pages. Recall from Section 7.3 that each step of a FREDO evaluation plan ultimately concludes by directing the subject to the “next” URL, so that the system may determine the next step for the subject in the protocol. Since the FREDO pages under discussion in this section are meant to function as steps within an evaluation plan, it stands to reason that each page will need to have some sort of link, button, or other mechanism for subjects to click on in order to indicate that they are ready to proceed to the next step.

As previously mentioned, the FREDO “next” url, requires certain data in order to correctly direct the subject. Most importantly, it needs to know the subject’s internal FREDO identification number. Authors of FREDO pages must be able to easily construct links, buttons, and other user interface elements that point to the correct “next” URL, and also contain the necessary data. Static page authors may do this manually, by making use of the various contextual data exposed by

FREDO. However, since this particular task is quite common (after all, virtually all static pages must include a “next” link), FREDO also provides a utility function (`next_link()`) to automate the process.

Pages are currently loaded into FREDO using a very simple web-based user interface. Figure 7.12 shows a screenshot of the page administration screen, after loading the page described in Listing 7.8 into the system; Figure 7.13 shows the final rendered appearance of this page in the subject’s browser. Figure 7.14 shows a real-life screenshot of a static page from an actual FREDO evaluation, and includes several notable user interface elements circled in red. At the top of the page, note the “progress bar,” which gives subjects an idea of how far along they are in in their evaluation plan. This feature is by default automatically inserted into every FREDO screen (including both static pages and surveys) seen by subjects. The second circled region is the link to page’s “next” URL; note that the link’s text can be set to whatever the page designer wishes.

While the FREDO page system is unquestionably basic, it is more than adequate for many of the data-presentation needs of an electronically-conducted study, particularly those that need to be able to support subjects speaking multiple languages.

7.6 External Instruments

FREDO’s intention is to act as a framework on which evaluations may be built. While there are certain common elements that many user studies share (questionnaires, static or semi-static information pages, etc.), there is obviously a great deal of diversity in terms of a specific evaluation’s data collection needs. As such, one of the most important features of FREDO is that arbitrary external data collection tools may be incorporated into evaluation plans.

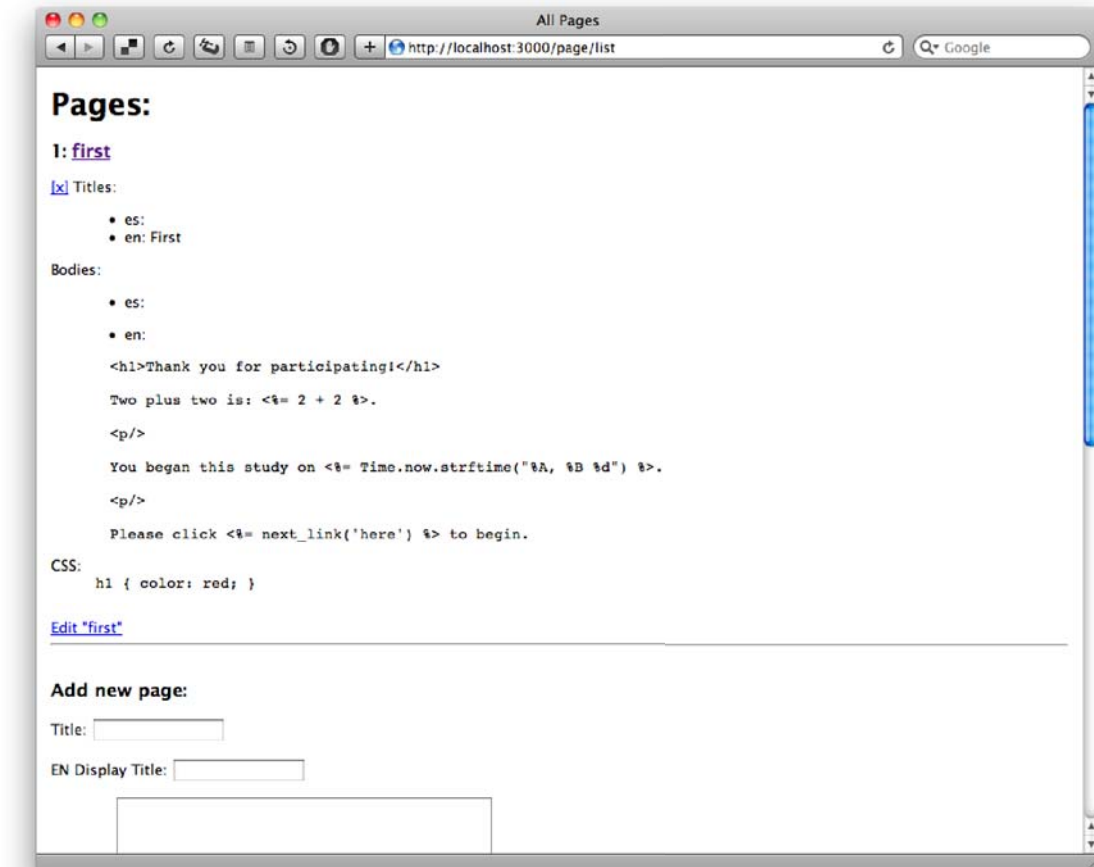


Figure 7.12: The page administration user interface. Existing pages appear at the top of the screen, and may be edited in place; new pages may be added using the controls at the bottom.

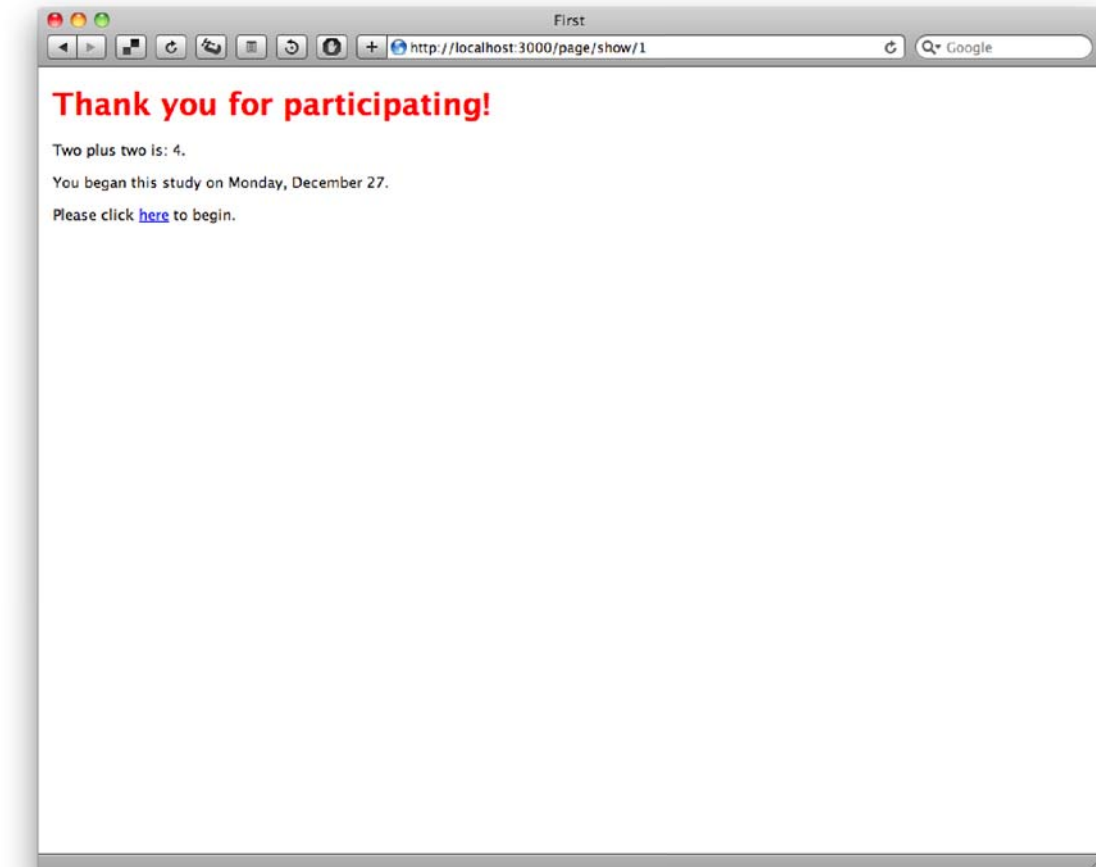


Figure 7.13: The page listed in Figure 7.12, in its final rendered form. Note that the dynamic elements (e.g., `<%= 2 + 2 %>`) have been computed and inserted into the final page view.

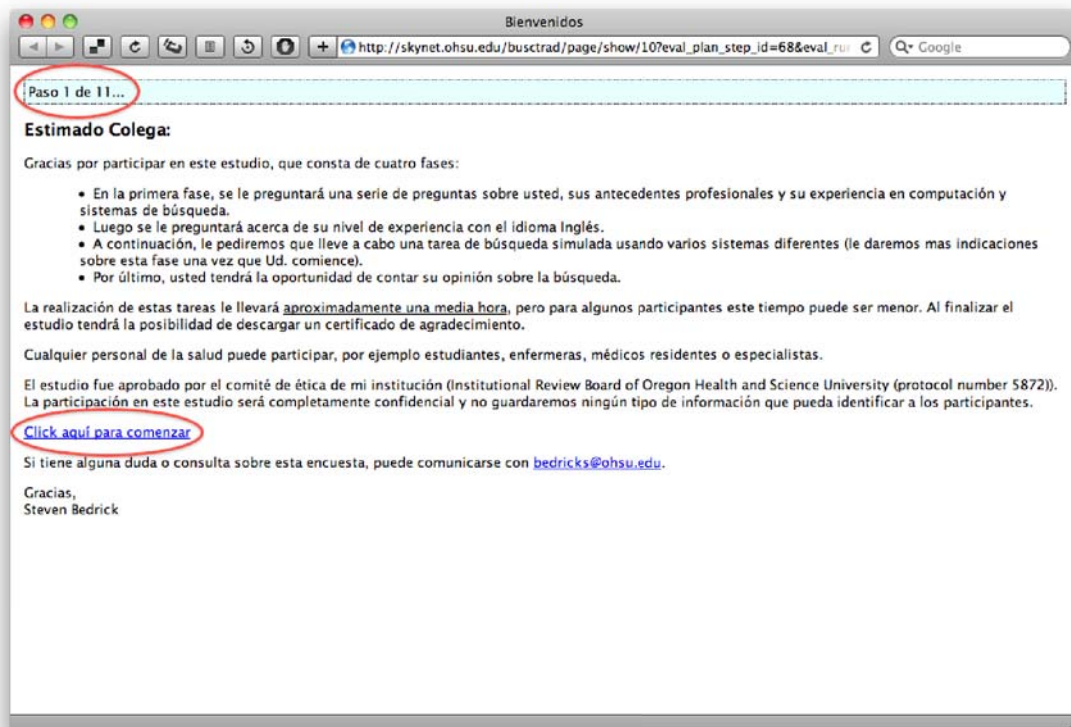


Figure 7.14: A real-life example of a page used during a study. Note the customized “next” link text, the Spanish locale, and the “progress bar” at the top of the screen.

As with virtually everything else involved with FREDO, this integration builds upon the HTTP protocol and employs standard patterns for tying disparate websites together. The simple evaluation plan shown in Listing 7.1 (on page 102) illustrates the concept: evaluation plan authors specify that a particular step is an “external” step, and provide the URL that they wish subjects to be directed to. They may optionally provide a set of key-value pairs that will be provided to the receiving URL.

When a subject is at a point in an evaluation plan where they must visit an external site, the system will send them there using a standard HTTP redirect.⁵ Along with the subject herself, FREDO will also send contextual information such as the subject identifier, the collector identifier, and so on, in the form of additional HTTP parameters.

This means that investigators wishing to integrate their data collection systems with FREDO need simply provide a URL endpoint to which subjects may be sent, and modify their systems to record whatever subset of the contextual data they deem necessary. The authors of a completely stand-alone data collection instrument, that did not need to link its data with that collected by FREDO questionnaires, might choose not to record any of the contextual information.

On the other hand, an instrument designed to be one part of a larger FREDO-managed experimental protocol might need to record the subject identifiers, so that individual subjects’ data may be linked with data collected by surveys or other external instruments. If the investigators knew ahead of time that their protocol would be using several collectors, they might choose to also record the collector identifier (so that they could easily tell which subjects had come from which sources).

Whether or not the external instrument explicitly *saves* the subject ID, it is im-

⁵As discussed in Section 7.3 on page 73.

portant that it retains the ID at least temporarily in some way. Ultimately, the external instrument will need to send the subject back to the FREDO system (see Section 7.3 and Figure 7.5), and, as previously discussed, the system depends on being able to identify which subject is returning from an external instrument.

In order to facilitate this return of control, FREDO provides external instruments with one additional piece of contextual data that has not been previously discussed: a “callback URL,” representing the destination to which the external instrument must send the subject once they have completed whatever it is that the external instrument is designed to do and are ready to continue onward through the protocol. By providing the callback URL in this manner, a single external instrument may be used by multiple FREDO installations without any special configuration on the part of its administrator.

7.6.1 Example external URL step: BuscTrad document selection

As discussed previously (see Section 6.6), one of the BuscTrad system’s purposes is to serve as a platform for user interface evaluation, and to this end may be loaded with “canned” article sets which may then be used as test content for any of the various result presentation interfaces described in Section 6.5. To support the evaluation described in chapter 8 of this dissertation, we added support for basic document selection to BuscTrad’s result views, and did so in such a way as to enable it to serve as an external instrument in a FREDO-managed study. This involved adding a URL endpoint to BuscTrad’s routing rules, and adapting some of the existing logic to check for the presence or absence of the FREDO-supplied contextual fields in each request. The process was very nearly trivial for an experienced web programmer to perform.

“Document selection” is a common feature of interactive information retrieval studies.[139, 140, 117, 126, 113] Oard defines document selection as a task in which

users are “given the documents that are nominated by the system as being of possible interest, the searcher must recognize which documents are truly of interest.” In other words, the task is meant to simulate a certain part of the search process: that of identifying and selecting the documents that a user considers to be relevant within the context of the experiment. By observing which articles a user selects, and comparing those to some ground truth, the investigator may objectively measure the user’s performance on the task.

The document selection interface that we constructed for BuscTrad looks almost exactly like the standard BuscTrad result presentation screens, with one small difference: the document selection screen contains no search controls (query input box, results-per-page, etc.), and instead simply lists the entries in a single article set. Additionally, the page as viewed by the subjects has embedded in it the subject identifier provided by FREDO, as described above. Another minor difference between the document selection interface and the standard search results interface is that, when viewing the document selection interface, the translated content comes from the pre-loaded translations (as described in Section 6.6) rather than being dynamically populated from Google Translate.

Each result, in addition to the “PubMed” and “Abstract” links, also has a link labeled “Mark.” When clicked, this link will cause BuscTrad to record that that particular article was clicked by the subject identified by FREDO, along with a timestamp and other metadata about the state of the interface (the language mode currently in use, etc.). The interface captures similar information when a subject clicks the “Abstract” link (to show or hide an article’s abstract). The page also features a hyperlink that subjects may click when they are finished reviewing an article set. Clicking this link will cause BuscTrad to send them back to the callback URL that FREDO provided (as described above).

The data captured by this interface are stored in BuscTrad’s relational database,

and are therefore easily extracted using SQL queries. Alternatively, BuscTrad features an interface designed for investigators to use to observe individual subjects' performance (see Figure 7.15). This interface allows investigators to review the behavior of a subject performing the document collection task. Recall that pre-loaded article sets can be created by combining two other sets. Often, this is done so as to create a set that includes both "relevant" and "not relevant" articles (i.e., one of the set's parents is comprised entirely of "relevant" articles, the other entirely of "not-relevant"). When using this interface to review the selections made by subjects viewing these compound sets, the investigator may specify one or the other of the "parent" sets as the "relevant" one, and the interface will thereafter treat articles originally from the specified set as relevant.

Investigators may view the specific articles that the subject selected, along with their relevance status, and may also download these data in CSV format. The interface also include a two-by-two table comparing the subject's document selections to the specified ground truth, and also automatically calculates several basic performance metrics (recall, precision, etc.) and so on.

In the specific case of BuscTrad, we have also prepared separate computer programs to perform bulk extraction of subject selection results. The administrative interface is primarily intended for use in small-scale studies or as a convenience tool for observing larger studies in-progress.

7.7 Future Development

While FREDO in its current form has proved to be quite useful, there are several additional features that would be useful. Evaluation plans currently do not support nested randomization/Latin Square blocks, both of which are necessary for certain experimental designs. Likewise, the current Latin Square implementation

Results: : Results for eval run 294

http://skynet.ohsu.edu/busctrad/results/eval_run_details?eval_plan

Evaluation Run: 294

Current filter:

None filter

Article Set: Final Merged set, 25% Diag. 75% Therap., random split, part 4 (id: 42)

Start time: 2010-09-24 02:59:38.463498
End time: 2010-09-24 03:03:23.863659

Includes articles from: TBI Diagnosis (final) (id: 20), TBI Therapy (final) (id: 21)

Relevant: TBI Diagnosis (final) Set Relevant

Ground Truth

	+	-	
User +	4.0	13.0	17.0
-	4.0	19.0	23.0
	8.0	32.0	40.0

Status:

- Recall: 0.5
- Precision: 0.235294117647059
- Specificity: 0.59375
- F-Measure: 0.32
- Accuracy: 0.575

[Toggle Details](#), [Toggle CSV](#)

ID	Parent Sub-Set	Relevant	Marked
Validity of the Patient Health Questionnaire-9 in assessing depression following traumatic brain injury.	Final Merged set, 25% Diag. 75% Therap.	1	t (TP)
Use of digital camera imaging of eye fundus for telemedicine in children suspected of abusive head injury.	Final Merged set, 25% Diag. 75%	1	f (FN)

2 errors occurred in opening the page. For more information, choose Window > Activity.

Figure 7.15: The main administrative screen for BuscTrad’s document selection interface, showing the selections made by one subject. The screen not only shows which article set the subject was viewing, but also which specific selections they made, along with the “ground truth” relevance of each document (if specified).

is relatively basic, and investigators have very little control over how it is configured. Adding additional configuration commands to the evaluation plan syntax would be helpful in this regard.

Furthermore, evaluation plans may not include branching logic, and steps occur entirely in isolation from one another. Adding a means for steps to pass subject-derived data between themselves would allow for more sophisticated evaluation plan steps, and would also enable conditional branching within an evaluation plan. As an example of what this functionality could allow, consider the case of a protocol in which a subject's response to a question in a survey would determine which of two different possible data collection instruments they should be assigned to.

Regarding surveys specifically, there are several obvious areas for enhancement. First and foremost, currently all fields are optional (i.e., subjects may choose whether or not to answer each field, and their choice will not affect their ability to complete the survey and continue in the protocol). Being able to designate certain fields as mandatory would be helpful in a number of situations. Similarly, some survey designs require that certain fields only accept data that meets a particular formatting requirement: only numeric characters, only valid email addresses, dates before or after a certain cutoff, etc. The FREDO survey system does not currently support any input validation rules, but future revisions will certainly include such features.

Another area for further development is in the question dependency logic. As currently implemented, the system is designed to support question dependencies between questions whose answers are well-structured and coded numerically (e.g., the various multiple-choice options or the drop-down menus). It shows or hides dependent questions based on a simple test of equality between a specified value and that selected by the subject (i.e., "Does the subject's response to question

X equal Y?”). This means that certain kinds of logical dependencies are impossible to encode into a FREDO survey.

For example, consider an author attempting to ask questions about age and voting. They wish to first ask subjects the question, “How old are you?” using a simple text box (allowing free-form entry by the subject). If subjects enter an age that is greater than or equal to 18, the author wishes to then ask the subject whether they are registered to vote.

Under the current system, it is not possible to build a survey with this exact form of question dependency. Instead, the survey designer would have to phrase the question as a binary choice (“Are you at least 18 years old?”) and link the voter-registration question to a “yes” response. Future improvements to the FREDO survey system, however, may ultimately allow survey authors to specify arbitrary dependency rules between questions. This is another area where FREDO’s syntax-based approach to survey specification may prove to be an advantage over more “user-friendly” tools, as the user interfaces for question dependency logic can be quite complex and cumbersome to use. Furthermore, menu-driven interfaces for specifying dependency logic are, by necessity, limited to whatever capabilities the original designer of the tool chose to include. Since the FREDO survey syntax is essentially a superset of the Ruby programming language,⁶ though, it should in theory be possible to allow authors to use any Ruby constructs they wish as part of their dependency rules.

⁶ Note that we chose the Ruby language for two reasons: first, it is a simple and easy-to-use language that manages to be both extremely flexible and also very easy to extend; and, second, the rest of FREDO is written in Ruby, so using it for the survey and evaluation plan syntaxes was the “path of least resistance.”

Chapter 8

Study Methodology

The main purpose behind translating search results from English to Spanish is to make it easier for Spanish-speakers to analyze and make use of those results. The evaluative component of this dissertation attempts to measure the degree to which the BuscTrad system (described in chapter 6) fulfills this purpose, and to identify under which of the system's user interface modes users perform best¹. Here, "perform best" could refer to a number of different aspects of user behavior: fastest, most accurate, most useful, etc. For the purposes of this evaluation, we looked primarily at a definition of "perform best" that considers the system's purpose: improving end-users' ability to make sense of and use English-language medical and scientific publications. A user who can read and understand a set of search results in some way is "performing well," and if they are better able to do so under user interface condition "A" than they are under condition "B", they can be said to be "performing better" under condition "A" than "B". The question, of course, is how to measure this.

One straightforward way of doing this would be to directly measure end-user

¹ Note that the various user interface modes are quite similar, and vary only in the languages used to present search results. This is a simplified and constrained setting, designed to isolate insofar as is possible the effects of language presentation from the myriad other factors that could affect user behavior or performance.

reading comprehension. A user who is truly able to make sense of and use a set of search results ought to have higher levels of reading comprehension with regard to those results than a user who is less able to make sense of the results. *Directly* measuring reading comprehension, however, is both time-consuming and difficult (though certainly not impossible). Instead, this evaluation attempted to *indirectly* measure end-user reading comprehension by using relevance judgments as a proxy for reading comprehension. To do this, we subjected users to a task whose successful completion depended on the user being able to understand each article's title and abstract well enough to make a judgment about the article's relevance to a pre-specified topic. This approach has the added benefit of being a relatively realistic simulation of the "results analysis" step of the information-seeking process[68].

Figure 8.1 (page 142) illustrates the flow of the evaluation. There are three main phases to the protocol:

1. Pre-evaluation questionnaires
2. Document selection task series
3. Post-evaluation qualitative feedback

This chapter discusses each of these phases in more detail. The OHSU Institutional Review Board reviewed this protocol (IRB # 5872), and in March, 2010 determined that it was exempt from approval requirements, since our protocol did not collect personally-identifiable information about any of our subjects. We used the FREDO system (described in chapter 7) to host and manage the protocol; the complete FREDO syntax describing the final evaluation plan may be found in Appendix A.

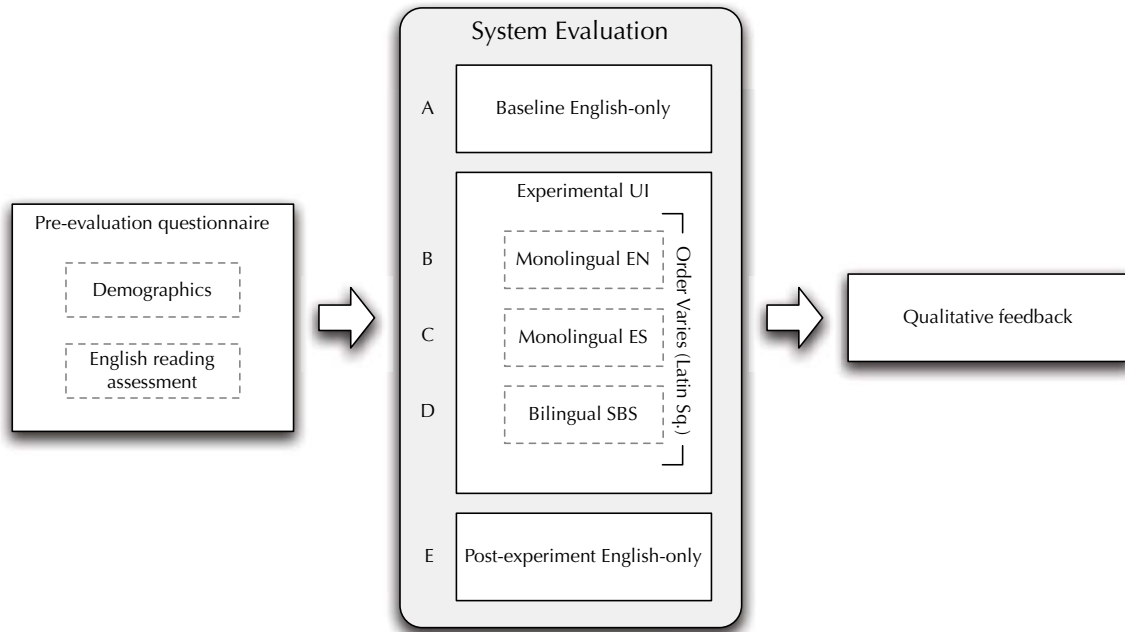


Figure 8.1: The overall flow of the evaluation protocol.

8.1 Survey instruments

The first phase of the evaluation consisted of a pair of survey instruments designed to collect demographic and other information about our subjects. The first instrument collected general data about our subjects, specifically their personal and professional demographics. The second instrument focused on their experience with languages, specifically but not exclusively English. Both included questions about many potential covariates or confounding factors to our outcomes of interest (see Section 8.3), and both included questions about aspects of professional life (tools and resources used, etc.) that to our knowledge have never been systematically collected from Latin American clinicians.

The study protocol included a third questionnaire, which subjects completed immediately after finishing the series of document collection tasks. This questionnaire solicited feedback from subjects about their experience using the BuscTrad interfaces, and also contained some open-ended questions that subjects used to

provide us, the investigators, with less strongly structured feedback.

All survey instruments were originally written in English, but were translated into Spanish with the aid of a native speaker. We used the FREDO system (described in chapter 7) to encode and present the three instruments.

8.1.1 Subject Demographic Questionnaire

The first survey instrument collected personal and professional demographic information from the subject. It began with basic questions (age, gender, nationality, etc.), and then asked a series of questions about the subject's self-perceived level of technical ability. This question set included a simple five-point Likert scale question about their level of computer expertise ("On a scale of one to five, how would you rate your computer skills?" with anchors running from "Unable to use a computer" to "Expert computer user"), questions about how often they use computers for personal purposes, what sorts of computer applications they regularly use, and so forth.

The next set of questions asked about the subjects' level of expertise at accessing data using internet search engines and other electronic resources, as well as whether or not they make use of social networking websites such as Facebook and Orkut. The purpose of these questions was to allow us to categorize our subjects' level of computer ability and experience.

Following these questions, the questionnaire moved on to questions about the subjects' professional demographics, starting with their medical role (attending physician, resident, student, etc.) and including questions about how many years they had been in that role. The questionnaire also collected data about the subjects' medical specialities. This section of the questionnaire also asked the subjects' levels of expertise regarding traumatic brain injury (see Section 8.2 for more information about why this is relevant), and also about subjects' *professional* computer use, in-

cluding which specific electronic resources (MEDLINE, etc.) the subjects used on a regular basis. Appendix B contains the complete text of the English-language version of this instrument; Appendix C contains the Spanish-language text.

8.1.2 Language assessment instruments

One major potential confounding factor in our study design is our subjects' language ability. Subjects with stronger English will likely experience the document selection task differently than subjects with weaker English, and it therefore seems probable that their opinions of the various BuscTrad interface modes could differ as well. This has traditionally been a stumbling block for cross-language information retrieval evaluation. Out of necessity borne from paucity of primary data, studies typically rely on self-reported information and somewhat rickety measurement scales. For example, the FlickLing system, described by Peinado et al. [155, 196], asks its users to report their level of ability for several languages on a different three-point scale: "active," "passive," and "unknown." While this is certainly better than nothing, it is a very coarse scale and could be missing important information.

However, the FlickLing scale is linguistically superior to the simplistic scales used by many other studies in that it at least recognizes the fact that language ability is a multidimensional construct. It is quite common, for example, for an individual to have different levels of ability with different aspects of a language (speaking vs. reading, for example), and the FlickLing scale makes an attempt to capture this information. Often, however, one sees even coarser scales that treat language ability as a unidimensional question. For example, Vasconcelos et al. used data from a Brazilian government database of researcher CVs, in which researchers rated their English proficiency on a three-point scale of "good," "reason-

able," and "poor"[39]. Again, this is better than nothing, but is also suboptimal².

For this study, we wished to measure our subjects' English reading proficiency levels, and we wanted to do so with more granularity and validity than either of the above-mentioned scales. We shared some of their constraints, however: formally testing English reading proficiency is time consuming and expensive to do in a valid way, and we were limited to data that can be self-reported. Furthermore, while we ideally would have been able to assess multiple dimensions of our subjects' English proficiency, time constraints limited us to only measuring the most important data (in this case, our subjects' English reading proficiency). Self-assessment, while by no means a substitute for formal evaluation, has the potential to be a valid approach for language assessment[197], and for the purposes of this study represented a useful middle ground between the oversimplified self-report scales described above.

The Interagency Language Roundtable (ILR), a US government body responsible for supervising language testing among federal employees and contractors, maintains a set of exhaustively-defined skill level descriptions for all aspects of language proficiency.³ Their scales are intended to be used to calibrate different testing instruments with one another, and describe five detailed "bands" of English proficiency in reading, speaking, writing, listening, and translation performance. The ILR scale also describes intermediate gradations between bands ("plus"-levels) to account for individuals who fall between one level and another.

The ILR has also produced self-assessment instruments which individuals may use to obtain a rough estimate of their overall level of proficiency. While these instruments are, of course, no substitute for a complete language assessment test, they may be adequate for our purposes. Therefore, for this protocol, we used the

²It should be noted that Vasconcelos and her colleagues did not *choose* this scale; it was a matter of making the best possible use of already-existing data.

³See <http://www.govtilr.org/> for complete descriptions of all aspects of the scale.

ILR reading self-assessment instrument. Appendix D includes the instrument itself, which is comprised of a series of statements about English proficiency with which subjects must indicate their agreement or disagreement.

The instrument consists of 21 statements about English reading proficiency, ranging from very simple (“I can understand simple instructions, such as in very straightforward street directions.”) to more complex (“I can understand both the meaning and the intent of writers’ uses of idioms, cultural references, word play, sarcasm, and irony in even highly abstract and culturally ‘loaded’ texts.”). They are arranged in increasing order of difficulty, and each is associated with a particular ILR band. Band-level scoring is done by identifying the highest level at which the subject agreed with *every* statement; Appendix F contains the Mathematica function used to calculate subject ILR scores.

In addition to the five primary ILR bands, the ILR scale also includes the concept of a “plus” level— i.e., a way to indicate that a subject’s score fell somewhere between bands. It is scored using a similar methodology to the main band scores: if a subject answers more than half of the next highest band’s questions in the affirmative, they may fall into a “plus” level. For example, questions three through five describe the first band level (designated as R-1 on the ILR reading scale), while questions six through ten describe the second band. If a subject agreed with each of questions three through five, but did not agree completely with questions six through ten, they would fall into the first ILR band according to the self-assessment’s scoring rubric. If, however they agreed with *at least half* of questions six through ten, they would score a “plus” designation.

The ILR self-assessment instrument is written in English. For this study, a native Spanish speaker translated the various statements into Spanish, and we then included the instrument’s statements as part of a larger survey instrument about our subjects’ use of the English language. In addition to the formal ILR

self-assessment questions, this instrument included several other questions about our subjects' use of the language in general, English in particular. The complete English-language text of the survey instrument may be found in Appendix G, and the Spanish translation may be found in Appendix H. Many of the questions in this instrument served multiple purposes. In addition to being potential covariates to our outcomes of interest, several were intended to provide some basic levels of convergent validity for the ILR instrument questions. Several other questions were included in order to provide a richer demographic picture of our subjects' interactions with English.

The first questions in the instrument asked the subject to report which languages they speak, both at home and in their professional context.⁴ The list of possible choices includes the major languages spoken in Latin America (Spanish, Portuguese, English, French) as well as several indigenous languages (Quechua, Aymara, etc.). Since we expected many of our subjects to be from Argentina, we also included Italian in the list of languages.

As discussed above, our primary method for measuring our subjects' English reading proficiency was the ILR self-assessment instrument. However, since (to our knowledge) no other IR studies had used this instrument, we decided to include simpler self-assessment questions in our survey instrument, both to serve as a "baseline" against which we might validate the ILR questions as well as a set of "backup" questions, in case the ILR instrument should prove unreliable or otherwise inappropriate.

The next block of questions was therefore intended as a set of naïve self-assessment questions similar to those typically used in language-related information retrieval user studies. We simply asked subjects to rate their own level of proficiency at reading, writing, speaking, and comprehending English on a scale of one through

⁴This pattern (asking the same questions about subjects' home and professional contexts) repeats itself throughout the instrument.

five, with one being low and five being high. The anchor text followed the pattern of “Cannot XXX any English/Perfect English XXX ability,” with the X’s replaced by reading/writing/speaking/listening and the remainder of the sentence arranged as was grammatically appropriate. We presented subjects with this series of questions twice; once for the general case of encountering English, and again for the case of encountering English in their professional capacities.

The next question asked subjects to rate their own level of overall English proficiency as compared to that of their colleagues (higher than average, about average, or lower than average). This question’s purpose was to act as an additional “sanity check” for the self-assessment questions, our hypothesis being that subjects who scored highly on the ILR instrument would in general report higher-than-average ability, and vice versa.

Following this question came the 21 ILR statements, as described above.

After the ILR statements, we presented subjects with several more questions about their background with, and daily use of, the English language. First, we asked how many years of formal English education they had undergone (another “sanity check” question), followed by a series of questions about the frequency with which they read/spoke/wrote/listened to English during their daily professional routines. Next, the instrument asked about the languages in which subjects used electronic resources to access medical information, as well as whether or not they had ever used machine translation tools such as Google Translate, Babelfish, etc., and (if applicable) the frequency with which they did so. Finally, the instrument concluded with an open-ended prompt inviting the subjects to share any additional thoughts they might have had about their relationship with the English language.

8.1.3 Post-task follow-up

After completing the series of document selection tasks, subjects completed a third survey instrument, containing a series of questions about their experiences using the various BuscTrad interfaces. The first group of questions sought to identify which interface (Monolingual English, Monolingual Spanish, or Bilingual) the subjects felt was most usable. The wording of the questions was inspired by the ISO 9241-11 conception of usability (“the extent to which a product can be used... to achieve specified goals with effectiveness, efficiency, and satisfaction...”)[198]. We chose the ISO definition as a starting point over other common definitions (e.g. that of Shneiderman and Plaisant[199]) because it was more amenable for use as a set of simple self-reportable questions (rather than a set of formal criteria or outcome measures for experimental evaluation).

The first five questions asked subjects to complete sentences of the general form “I thought that the XYZ system was the ABC one to use”, with “XYZ” replaced by a drop-down menu whose options were the three interface modes (written as “English-only,” “Spanish-only,” and “Bilingual”), and “ABC” with a dimension of usability (“easiest,” “most efficient,” “fastest,” “liked best,” and “most difficult”). The complete text of the English version of this instrument’s questions may be found in Appendix I, and the Spanish version of the text in Appendix J.

Note that there were a total of four “positive” questions and one “negative” question (“most difficult”). While the four positive questions are asking about different aspects of the subjects’ experiences, we did expect individual subjects’ responses to these questions to be *roughly* similar— i.e., if a subject indicated that they found the English-only interface to be the easiest one to use, we would be surprised if they then proceeded to indicate that they “liked” the Spanish-only interface the best. Similarly, we anticipated that individual subjects’ responses to the negatively-worded question would be *roughly* different from their responses to

the positively-worded questions.

Immediately following these single-choice questions was a series of free-text questions asking the subjects to report, for each of the three interfaces, what they found *easiest* and *most difficult*. The purpose of these questions was to give subjects an opportunity to leave open-ended feedback about their experiences using the various interface modes.

8.2 Document Selection Task

As described above, the main part of the experimental protocol consists of a series of modified document selection tasks, in which subjects are asked to review a set of documents and identify (“select”) a subset of the results according to some set of criteria[113, 126]. By holding constant the set of articles from which they are to choose, varying the way in which those articles are presented, and measuring various aspects of the subjects’ behavior during the task, we attempted to assess the degree to which presentation mode affected subject behavior.

In this case, the selection criterion was whether or not an article was, in the judgment of the subject, relevant to the topic of *traumatic brain injury diagnosis*⁵. We chose this criterion for several reasons. First and foremost, traumatic brain injury is a well-studied clinically-relevant topic (for example, see [200, 201]), with a rich body of literature from which to draw. Second, as described in Section 8.4, we knew that one of our major sources of subjects was going to be a network of physicians engaged in traumatic brain injury research, so the topic was one that would be of interest to many of our subjects.

The overall purpose here was to simulate the task of identifying relevant arti-

⁵The complete instruction given to subjects specified that they were to consider articles about “diagnosing, screening for, or otherwise measuring traumatic brain injury and its sequelae” as relevant.

cles from a set of search results. By comparing user relevance judgments under each condition with our ground-truth knowledge of result “relevance,” we can indirectly examine how presentation mode affects the subjects’ ability to review and evaluate search results.

For this evaluation, we used the BuscTrad document selection interface described in Section 7.6.1 to present a set of simulated results from a literature search about a particular medical topic, using the various result presentation modes described in Section 6.5.⁶ As mentioned above, the topic under consideration by our subjects was that of traumatic brain injury (see Section 8.2.1 for more details about the document set itself).

Using the tools for pre-loading (and combining) article sets into BuscTrad described in Section 6.6, these simulated search results were composed of two subsets of articles: one known *a priori* to be “diagnostic” in nature (i.e., articles about diagnosing traumatic brain injury), and another known *a priori* to be “therapeutic” in nature (i.e., about treating traumatic brain injury). See Section 8.2.1 for more details about the document set.

As shown in Figure 8.1, subjects performed this task a total of five times. The first and last times were under the monolingual English condition, and served as warmup and follow-up tasks. The three remaining sessions were divided among the three interface modes, such that subjects participated in each mode once. To control for the possibility of an order effect, the order in which subjects completed the three middle selection tasks was varied using a simple Latin Square, as described in [137] and illustrated in Table 8.1.

⁶Recall that BuscTrad supports three primary results presentation modes: monolingual English, monolingual Spanish, and a bilingual mode that presents English and Spanish results side by side.

1	2	3
Mono. EN	Mono. ES	Bilingual
Bilingual	Mono. EN	Mono. ES
Mono. ES	Bilingual	Mono. EN

Table 8.1: The latin square used to balance the order in which subjects completed the selection tasks.

8.2.1 Document collection

As mentioned above, the documents that subjects reviewed while completing the selection task were articles from the medical literature about traumatic brain injury (TBI). For the purposes of this study, articles about *diagnosing* TBI were used as the “relevant” documents, and documents about *treating* TBI (“therapies,” etc.) were categorized as “non-relevant.” The intention was to ensure that both the relevant and non-relevant articles were relatively similar with respect to their general subject matter, and differed only in their focus (i.e., on diagnosis vs. therapy). This would, in theory, force subjects to focus on the contents of the articles themselves in order to make accurate relevance judgments. We used diagnosis and therapy primarily because we felt that they represented two well-understood and easily-distinguishable dimensions of clinical reasoning, and also because their respective bodies of literature were well-separated from one another.

Given our time and personnel constraints, we wished to develop a semi-automatic methodology for building our document collection. To do this, we combined combine two “pure” sets of articles: one about TBI diagnosis, and another about TBI therapy. We built these pure sets using PubMed’s “Clinical Queries,” which are based on the work of Haynes, et al.[202] and allow users to focus their queries on particular clinical sub-topics (therapy, etiology, prognosis, etc.).[203] We further limited both sets to include only articles that relate to TBI diagnosis or therapy *in humans*. We generated our two pure sets by using BuscTrad’s article set function-

ality (described in Section 6.6) and the following PubMed queries:

1. (traumatic brain injury) AND (Diagnosis/Narrow[filter]) AND Humans [mh]
2. (traumatic brain injury) AND (Therapy/Narrow[filter]) AND Humans [mh]

We collected the first 200 results from each query, sorted by publication date. For the purposes of this study, an important attribute of these two sets was that they be completely disjoint (i.e., no articles from the diagnosis set may be present in the therapy set, and vice versa). When we encountered duplicate results in the two sets, the duplicates were removed from both and replaced with the next result in line from the original raw search results. We found that result sets from the “diagnosis” and “therapy” Clinical Query filters overlapped much less than other combinations (e.g., “diagnosis” and “prognosis,” etc.), further validating our choice of diagnosis/therapy as the relevant/not-relevant criteria.

Once we had assembled satisfactorily pure diagnosis and therapy article sets, we used BuscTrad’s article-set manipulation tools to randomly select 50 (25%) from the “diagnosis” set, and 150 (75%) from the “therapy” set, and then combined the two sub-sets. This gave us a 200-article meta-set, each article of which was “known” to be both a) about TBI in humans, and b) either diagnostic or therapeutic in nature. Next, we randomly split this meta-set into five “piles” of 40 articles, to be used during the five evaluation steps described earlier in this chapter and illustrated in Figure 8.1. Figure 8.2 illustrates this splitting process. By splitting randomly, we ensured that the piles would be relatively homogeneous with respect to diagnosis/therapy ratio, publication date range, and so on.

Five article piles and five document selection task steps equals twenty-five possible combinations of article pile and selection task. Since we were unsure of how

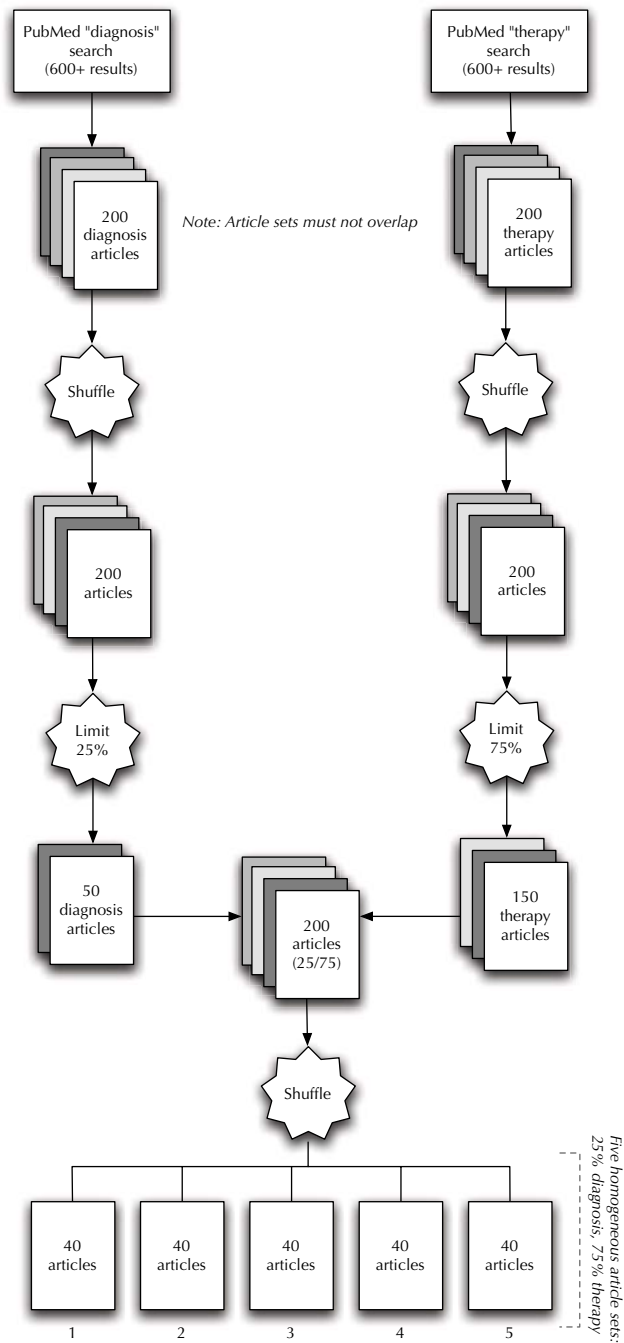


Figure 8.2: Document set creation, from the initial single-domain result sets, to the combined meta-set, to the five “piles” or sub-sets of the meta set.

many subjects we would be able to recruit, however, we were leery of randomly assigning article piles to tasks on a per-subject basis: if we only ended up with thirty or forty subjects, some article pile/selection task combination might (by chance) go unused, or, worse, end up over-represented, thereby introducing difficult-to-characterize bias into our evaluation. To avoid this, we randomly assigned each article pile to one evaluation step, and kept that assignment constant across subjects. Additionally, we randomized the order in which subjects interacted with each pile/mode combination.

While this solution was far from perfect, and opened the door to biases of its own,⁷ we felt that it represented the best compromise, and that the randomization (both in terms of the piles' contents, as well as their assignment) helped to minimize the possible biases.

Initial validation

This primarily automated approach to document collection construction and curation is highly experimental, and before using it we sought to validate the accuracy of the Clinical Query-derived "relevance judgments." The central question we wished to address was whether human subject experts would, upon reviewing a set of articles, agree substantially with the class assignments ("diagnosis" and "therapy") made by the Clinical Queries.

To accomplish this, we enlisted the help of three clinical researchers with expertise in traumatic brain injury and deep familiarity with the field's body of literature. We asked them to use an early prototype of the BuscTrad document selection interface to review a random selection of 100 articles, half of which were, according to the Clinical Queries, about diagnosis of TBI, and half of which were about treating TBI. Their task was to identify articles that were about diagnosis.

⁷For example, if the articles in the pile assigned to the Monolingual Spanish interface mode were particularly poorly translated as compared with those assigned to the Bilingual interface mode...

Initially, the raters exhibited low levels of agreement with one another and with the Clinical Queries. In discussing the task with the raters, we realized that, being expert systematic reviewers, they were using very narrow and technical definitions of both “diagnosis” and “therapy,” and as such identified very few articles as being of either category. This series of discussions led us to formulate more precise instructions to the raters, which we ultimately ended up using for the final study itself (see Section 8.2). These instructions included a definition of “diagnosis” that was more in keeping with how we anticipated that biomedical professionals who were *not* expert systematic reviewers might conceive of the term.

Using the revised selection criteria, the raters achieved much higher levels of agreement, both between themselves and with respect to the Clinical Queries. Within themselves, the Fleiss’ Kappa between the three raters was 0.793, well into the range generally considered to represent “substantial agreement.”[204] The Cohen’s Kappa scores between the three raters and the Clinical Queries were 0.774, 0.752, and 0.731. Based on these scores, we decided that the diagnosis and therapy Clinical Query filters were sufficiently accurate to use as the basis for this study’s document collection.

8.3 Hypotheses & Outcome Measures

This evaluation sought to address three main hypotheses:

1. Subjects would *prefer* BuscTrad’s bilingual interface mode to its two monolingual modes;
2. Subjects *perform better* (i.e., more accurately) at the document selection task when using the bilingual mode than when using the two monolingual modes;
3. Subjects would be able to complete the document selection task *faster* under

the monolingual Spanish mode than under the bilingual or monolingual English modes.

Furthermore, we hypothesized that individual subjects' preference and performance would depend in part upon their English proficiency. We anticipated that subjects with stronger English ability would experience less of a benefit from using the translated results than would subjects with weaker English skills,⁸ and so would therefore be less likely to prefer the translated results (or the interfaces in which they are used).

For the first hypothesis, our outcomes of interest were subjects' responses to the feedback questions in the third survey instrument, described in Section 8.1.3. Given that there were three possible responses to each question (one for each of the interface modes), there were several possible measures from which we could choose. One approach would simply be to examine the relative proportions of each reply; i.e., how many subjects responded to each question with each possible answer. For example, consider the question asking subjects which interface mode was easiest to use. We could simply look at each response in isolation (e.g., 30% of the subjects answered "English-only," 30% answered "Spanish-only," and 40% answered "Bilingual").

However, many common approaches to categorical data analysis work better with dichotomous (binary) data than with polychotomous data. For example, odds ratios and logistic regression both depend on at least one of the variables under consideration being binary. In order to make use of such analytical tools, we considered dichotomizing this trichotomous variable by, instead of simply looking at the reported proportions, *grouping* some possible answers together and creating binary *contrasts* between one response and the other two responses. For exam-

⁸ In fact, depending on the translation quality, it seemed to us to be possible that subjects with strong English skills would rather *use* those skills to read the original results rather than attempt to decipher what may be poorly-translated text into their mother tongue.

ple, we might look at the proportion of subjects who found the bilingual interface easiest as compared to those who found either of the two monolingual interfaces easiest. This, in effect, examines the effect of bilingual interfaces against that of monolingual interfaces.

The other two possible contrasts compare the number of subjects finding the English-only interface easiest to the number finding either the bilingual or Spanish-only interfaces easiest (“English-only” vs. “some Spanish”), and the number finding the Spanish-only interface easiest against the number finding either the bilingual or English-only interfaces easiest (“Spanish-only” vs. “some English”). We included these binary proportions in our set of metrics regarding hypothesis #1.

The second hypothesis depends on measuring subjects’ performance at the document selection task described in Section 8.2. We used several standard information retrieval performance metrics to analyze subject performance, specifically, *recall*, *precision*, and the *F-measure*. These metrics are typically used in the evaluation of information retrieval *systems*[194], but are also widely used in evaluating subjects’ performance at interactive information retrieval tasks[137]. All of these metrics are ultimately derived from the set of articles retrieved by a given system (or selected by a given subject) and from the set of articles belonging to the “relevant” set (in this case, articles about TBI diagnosis).

Recall represents the proportion of the relevant articles that the subject happened to select, and is equivalent to a screening instrument’s *sensitivity* or a classifier’s *true positive rate*:

$$recall = \frac{\# \text{ relevant selected}}{\text{total } \# \text{ relevant}} \quad (8.1)$$

Precision is analogous to a screening instrument’s *positive predictive value*, and represents the fraction of the selected (retrieved) articles that happened to be relevant:

$$precision = \frac{\# \text{ relevant selected}}{\text{total \# selected}} \quad (8.2)$$

Both metrics have shortcomings, and when looked at in isolation do not necessarily give an accurate picture of a subject's performance. For example, a subject may achieve perfect recall by simply selecting every possible article, and may similarly achieve perfect precision by selecting only one article (albeit one that happens to be relevant). Of course, in both cases, by optimizing for one metric, the subject would hurt their score on the other, and so the two metrics are typically reported together.

One commonly-used composite measure that combines precision and recall into a single value is Van Rijsbergen's *F measure*[205], which is commonly described as a weighted harmonic mean of precision and recall.⁹ For some evaluative purposes, it may be desirable to weight precision or recall differently. One classic example of such a situation is that of a search engine designed to retrieve patents. In this case, recall would be of paramount importance, since even a single missed result could turn out to have very costly repercussions. Therefore, when evaluating patent retrieval systems, we might wish to weight our metric in such a way so as to penalize decreased recall while rewarding increases in the same.

In most cases, however, users of the F measure weight precision (P) and recall (R) equally, in which case the measure is referred to as the *balanced F measure*, and is calculated thusly:

$$F = \frac{2 \times P \times R}{P + R} \quad (8.3)$$

The third hypothesis, that subjects would complete the document selection task most quickly under the monolingual Spanish mode, is the most difficult one to

⁹Consult [194] for a complete discussion of the F measure, including an explanation of why it is calculated using the *harmonic* rather than *arithmetic* mean of precision and recall.

evaluate given this study's protocol. For the purposes of our analysis, we measured this by simply taking the elapsed time (in seconds) between the subjects' first and last document selection within a given interface. Previous studies (for example, those of Karlgren and Hansen[162, 206]) have found that users are able to perform this sort of task more quickly when working in their native language than when working in a foreign language (even one in which they are quite fluent), and we expected this to be the case for the subjects in the present study . However, it is important here to note that this design contains some important limitations relating to how we may interpret speed-related data. With a reading-intensive task such as the one used in this study, time and accuracy are often inversely related. That is to say, the more time a subject spends on the task, the greater the accuracy they will display (as opposed to a subject who "rushed through" the reading task and therefore displayed reduced accuracy).

We may therefore think of each user as having a distinctive "time/accuracy" profile, relating perhaps to their level of reading comprehension, familiarity with the topic of the text under consideration, and so forth. However, we must also consider the possibility that the different interfaces (Monolingual English, Monolingual Spanish, and Bilingual) affect their users' time/accuracy profiles in different ways. In other words, the relationship between speed and accuracy may be affected by the interface mode (see Figure 8.3).

As described by Sperling and Doshier[207], varying experimental parameters (such as interface mode) may affect the speed/accuracy strategy exhibited by an individual subject. Our experimental design does not include controls for this phenomenon, and as such within-subject speed measurements may not be as informative as they might at first seem— it is impossible for us to know whether a difference in a subject's speed from one interface to another resulted from the interface change itself or from a difference in the subject's speed/accuracy tradeoff resulting

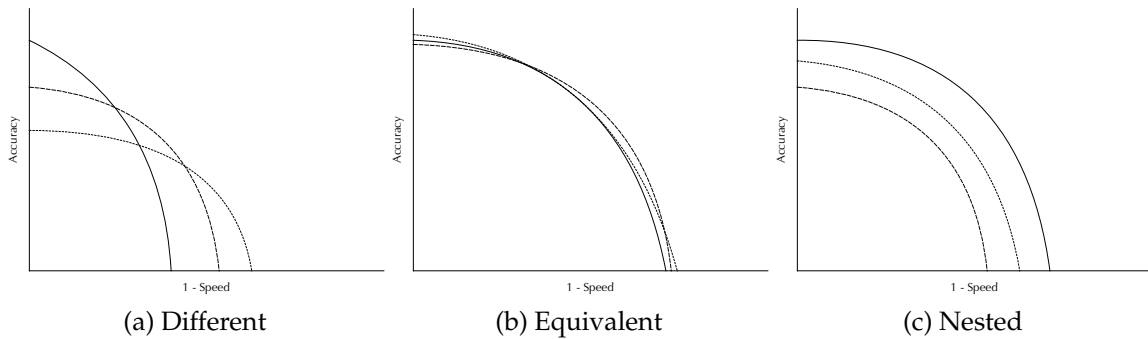


Figure 8.3: Three possible hypothetical performance operator characteristic (POC) combinations, illustrating possible strategy changes between interface modes. Each line represents a given user's potential POC under a different interface mode (monolingual English, etc.). See Sperling and Doshier[207] for a detailed discussion of this phenomenon.

from the interface change.

Because we are not enforcing any time limits on our subjects, different users may choose to spend more or less time on each task. This represents an uncontrolled source of variation, and will affect our ability to draw conclusions about the relationship between interface mode and speed. However, for our purposes, we felt that this simplistic approach would still yield some useful data, so long as the design's limitations (and therefore the limits to how these data may be interpreted) were taken into account when considering the results.

8.4 Subjects & recruitment

The subjects for this study were Latin American clinical professionals who work primarily in the Spanish language. All types of clinicians were eligible for participation, as were medical librarians, biostatisticians, informaticians, etc. Our goal was to recruit as diverse a population of subjects as possible. To this end, we used several recruitment strategies. First, we reached out directly to specific colleagues and acquaintances in Latin America, particularly including researchers at Hospi-

tal Italiano and Hospital Pedro de Elizalde in Buenos Aires, Argentina; Hospital de Emergencias Clemente Álvarez, in Rosario, Argentina; and MEGASALUD, in Santiago, Chile.

In addition to directly contacting specific individuals to assist us in subject recruitment, we made use of several newsgroups and practice networks, specifically including the International Collaborative Trauma and Injury Research Training network, and the IMeCA-LAC (Latin America & Caribbean) Google Group,¹⁰ which is an online discussion group frequented by numerous Latin American medical informatics professionals.

The third recruitment avenue that we pursued was the well-known “snowball” method. Our recruitment letter (see Appendix K) included a request for the reader to forward the letter to any of their colleagues that they felt would be interested in participating. While this approach was not as successful as we originally had hoped, we do know of several subjects who, after participating in the study, went on to attempt to recruit their friends and colleagues (with varying degrees of success). In the end, we created a total of 13 different collector URLs (see Section 7.2), each one of which represented a specific individual or group that was assisting us with subject recruitment.

For both budgetary and logistical reasons, we were unable to provide our subjects with monetary compensation for participating in our study. However, we were able to provide our subjects with certificates of completion upon finishing the the study. Figure 8.4 contains an example of what the certificates looked like. We based the wording on certificates used in other studies. Note that the certificate is in English; we had initially designed Spanish-language certificates, but changed them to English after consulting with Latin American colleagues.

We created a web-based form that subjects could use to generate personal-

¹⁰http://groups.google.com/group/imeca_lac/

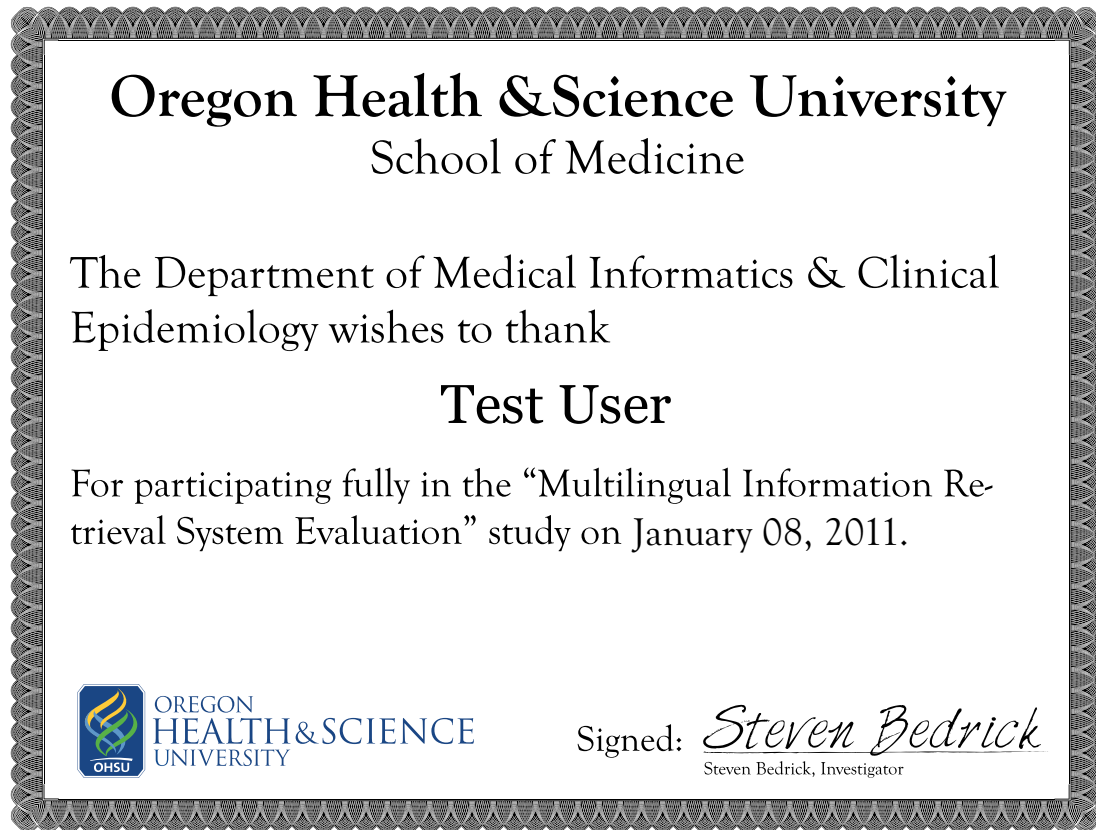


Figure 8.4: Subjects who completed the study protocol were able to generate and download personalized certificates of completion.

ized certificates. The certificates were in the easy-to-print Adobe PDF format, and prominently featured the Oregon Health and Science University logo. In accordance with the rules of our IRB exemption, we did not log or otherwise record the names subjects used for their certificates, although we did record subject identification numbers and timestamps for each certificate.

8.5 Pilot Testing

We performed two short rounds of pilot testing. The first round took place in November, 2009 at Hospital Luis Vernaza in Guayaquil, Ecuador, and involved early prototypes of the selection interface, document collection, and instructions

(see Figure 8.5). Our subjects for this test were approximately ten early-stage medical students, most of whom had little to no experience using bibliographic search tools such as PubMed. The medical students seemed to understand the task perfectly well, seemed to have no trouble identifying and selecting articles, and expressed appreciation of the translated search results. Unfortunately, since this was a very early pilot test, the system still had numerous technical flaws, one of which resulted in the system failing to properly record the subjects' selections. As such, we were unable to assess the performance of the subjects in this round of the pilot testing.

The second round of pilot testing took place in the summer of 2010, and was intended as "dry run" of the final versions of the system, instructions, and survey instruments. For this round, we enlisted the help of approximately ten staff members (including both doctors and nurses) in the critical care unit of Hospital de Emergencias "Dr. Clemente Alvarez" in Rosario, Argentina. Subjects used the same study materials as the rest of the study's subjects, but were personally recruited by a colleague in the unit. They were specifically asked to identify confusing or difficult questions and instructions, and their colleague was familiar with the study's aims and materials. All of these pilot subjects ultimately completed the study successfully, and no major issues arose, leading us to conclude that the system and study materials were ready for a full launch.



Figure 8.5: Pilot testing an early version of BuscTrad at Hospital Luis Vernaza in Guayaquil, Ecuador, November 2009.

Chapter 9

Results

The first subjects participated in the study in July of 2010, with data collection completed in early October. The study's landing page was viewed a total of 242 times via the 13 collector URLs (see Section 8.4), and 145 subjects at least "began" the study by proceeding past the landing page. Ultimately, a total of 77 subjects "completed" the protocol— meaning that 77 subjects proceeded through every step of the protocol, from the initial "landing page" through the final conclusion page. Of these 77, we censored 18 subjects for a variety of reasons, leaving us with a final total of 59 subjects. Three subjects simply "clicked through" the study's various screens without participating in any of the study activities (responding to questions in the survey instruments, selecting articles using the various interfaces, etc.).

Another three subjects participated in the first two survey instruments, made no document selections and did not participate in the final survey, but clicked through the remaining screens (and therefore technically "completed" the study, from FREDO's perspective). We have no way of knowing if these subjects made no selections because they found no articles to be relevant, or because they simply did not do the task; therefore, we removed them from consideration. The remaining 12

censored subjects either skipped one or more of the study's steps, or, in a very small number of cases, experienced technical issues preventing their data from being captured correctly by the system. Figure 9.1 provides a QUORUM diagram[208] describing the subject inclusion flow.

Figure 9.2 illustrates the number of subjects who left the study at each stage of the protocol. Note that out of the set of 145 subjects who actually began the study, 40 never made it as far as the document selection tasks, and did not progress beyond the initial survey instruments (see Section 10.4 for further discussion on this subject).

As previously described, the data for this analysis were collected using the FREDO system's survey tools (see Section 7.4) as well as BuscTrad's document selection interface 7.6.1. We analyzed the data using Mathematica 8.0,¹ SPSS 19,² and R 2.11[209].

9.1 Subject Demographics

As described in Section 8.1, an important aspect of this study was the collection of demographic information about our subjects. The survey instruments described previously covered a wide range of topics; here we report the subset of our subjects' responses that we believe to be most relevant to the remainder of the study. The complete text of the instruments (in both English and Spanish) can be found in appendices B–J. The results discussed here are for the portion of subjects who successfully completed the study (as described above); in a small number of cases, however, members of the set of censored subjects possessed interesting or otherwise notable demographic features that were not present in the set of included subjects. In those cases, we have used footnotes to present the extra data.

¹Wolfram Research, Champaign, IL

²IBM, Somers, NY

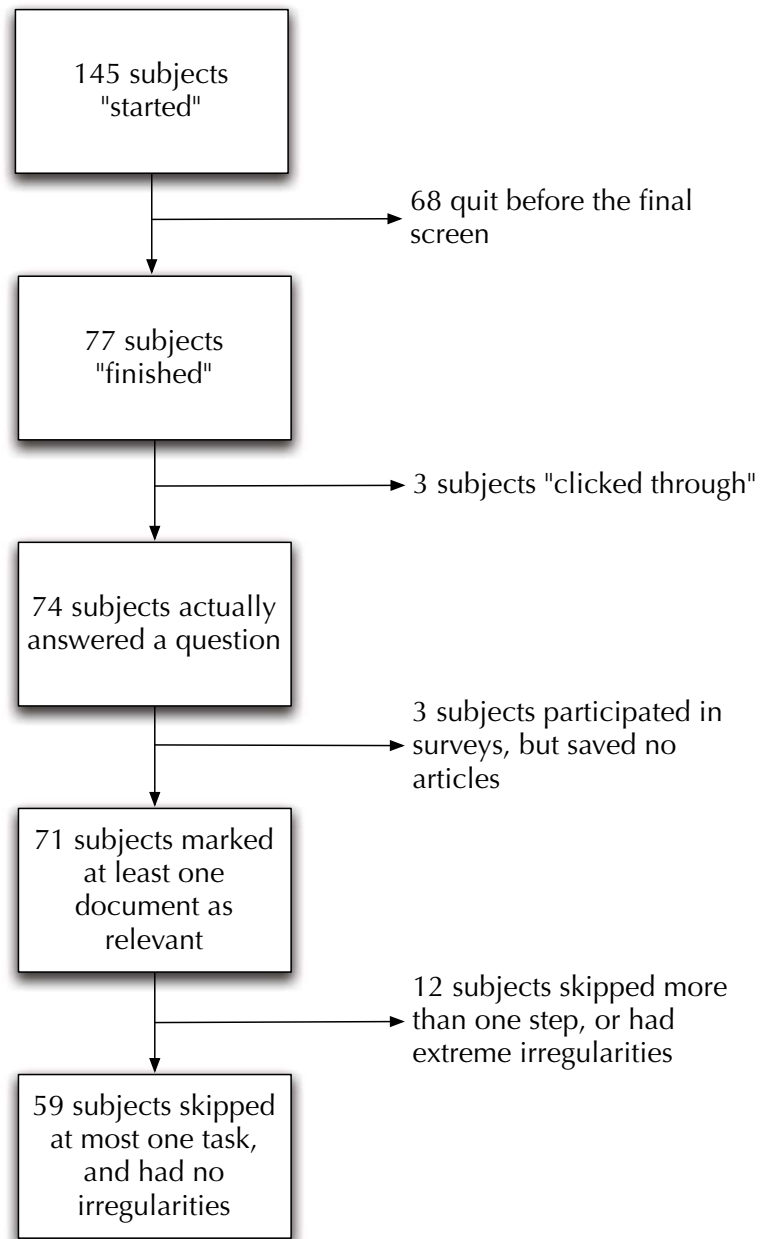


Figure 9.1: QUORUM diagram[208] showing the subject inclusion flow of this study.

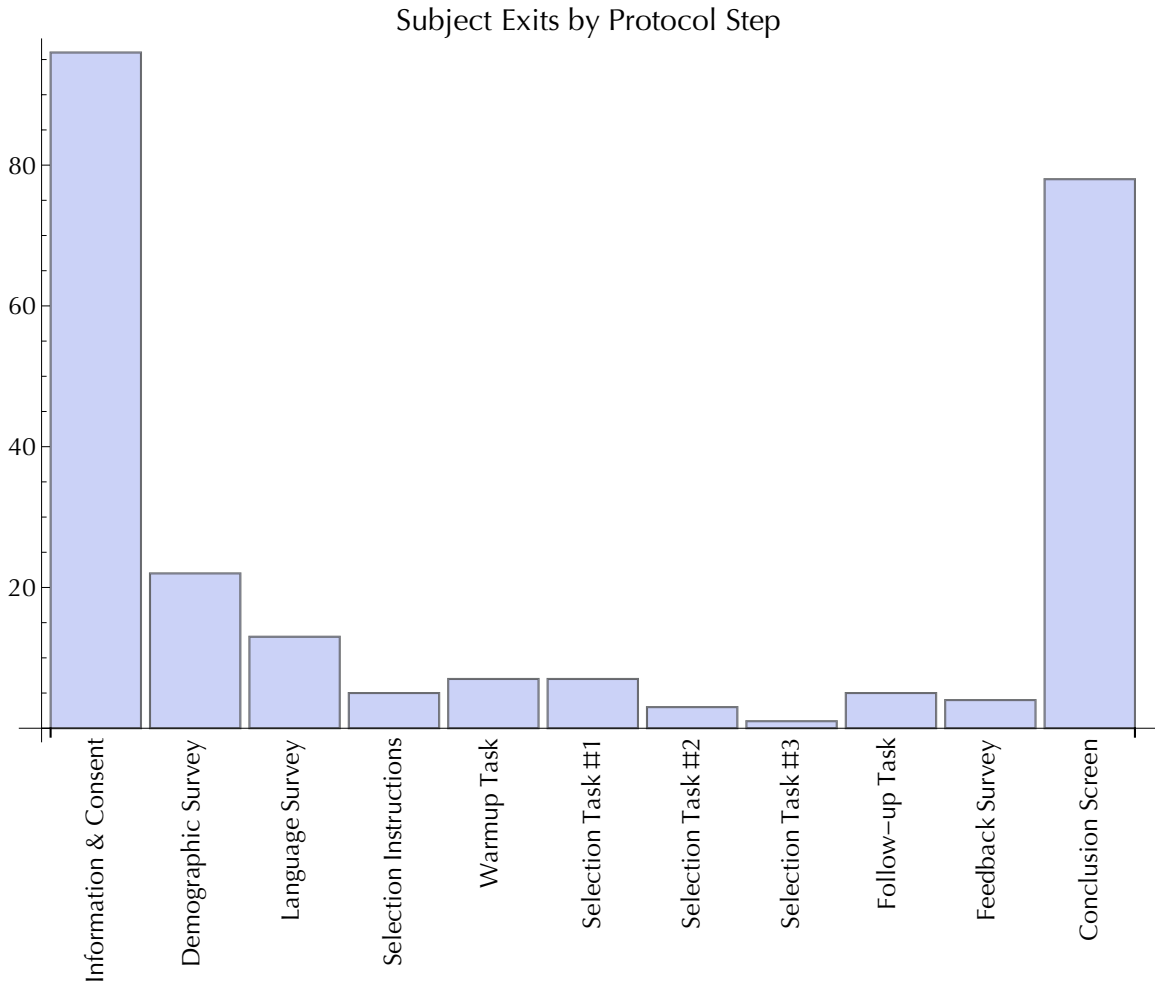


Figure 9.2: At each step, some number of subjects decided that they did not wish to complete the study. Each bar in this figure represents a step in the protocol, and the height of the bar represents the number of subjects for whom that step was the last one at which they interacted with the system.

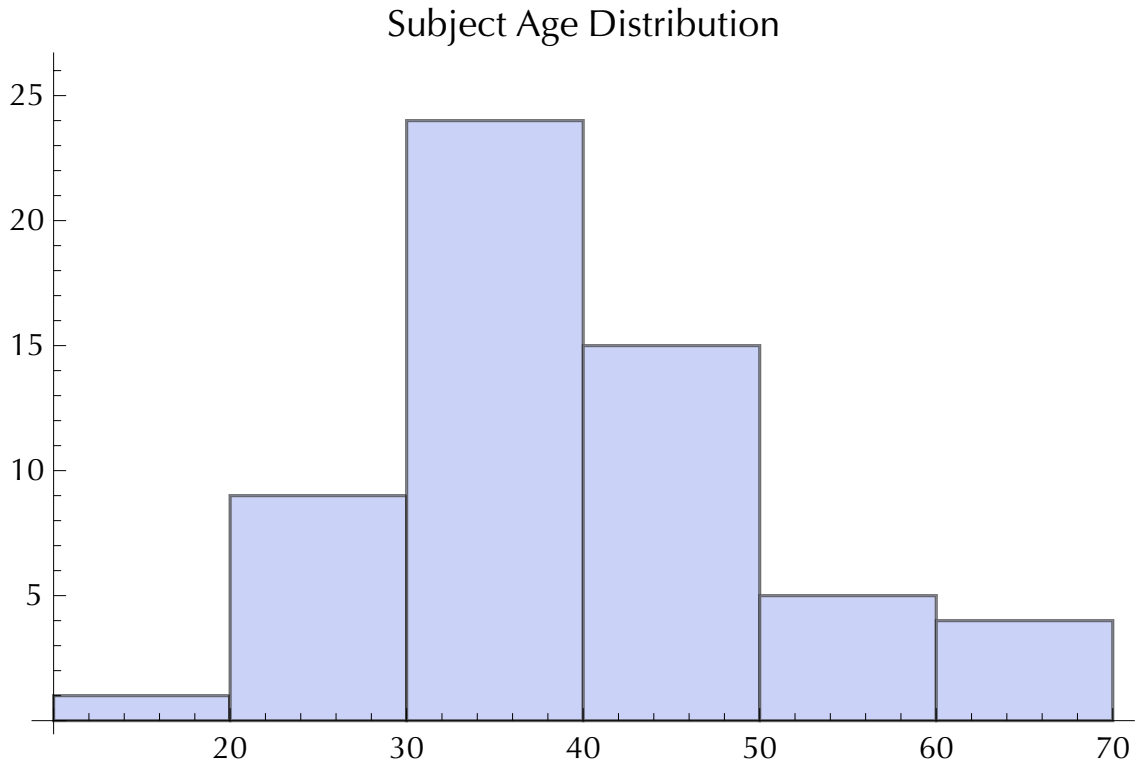


Figure 9.3: Age distribution of participating subjects.

General Characteristics

The 59 included subjects ranged in age from 19 to 63 years old, with a mean and standard deviation of 38.1 and 10.6 years, respectively. Figure 9.3 illustrates the age histogram; note that our subjects' ages, while slightly skewed towards the younger end of the spectrum, were far from exclusively so. There were more female subjects ($n = 33, 57\%$) than male subjects ($n = 25, 43\%$). The mean age for female subjects was slightly lower than that of male subjects (36.9 years for females as opposed to 39.8 for males; see Table 9.1); however, there was no statistical difference between the two groups (Independent-samples t -test, one-sided p -value: 0.16). Subjects reported working in eleven different countries (see Table 9.2), representing most of South and Central America (see Figure 9.4). Only six subjects reported working in countries other than that in which they were born; five of these were from one neighboring Latin American country to another.

Age	Male	Female
10–19	0	3
20–29	19	12
30–39	35	45
40–49	19	30
50–59	12	6
60–69	12	3

Table 9.1: Subject counts broken out by age and gender.

Country	Count
Argentina	26
Chile	12
Bolivia	5
Colombia	5
Ecuador	4
Venezuela	1
Uruguay	1
Peru	1
Guatemala	1
Nicaragua	1
Mexico	1

Table 9.2: Subject counts, by primary work country.



Figure 9.4: Subjects in this study worked in nearly every Latin American nation. Countries in which at least one subject reported working are shaded in blue; black countries contributed no subjects.

Medical Demographics

Recall from Section 8.4 that subjects for this study were all medical professionals of some kind. Subjects self-identified their medical role, and could choose from “Attending Physician,” “Resident,” “Nurse,” “Medical Student,” and “Other;” Table 9.3 details the numerical breakdown by role, and Table 9.4 shows the age distribution across roles. We provided no explicit definitions for these categories, and relied on our subjects’ expertise and ability to self-categorize. Of the named roles, attending physicians were most common, followed by residents, nurses, and medical students. 14% of subjects self-identified as “Other,” including two dentists, two medical informaticians, a medical librarian, a speech therapist, and a clinical researcher specializing in traumatic brain injury.

When asked to report their individual medical specialties, our subjects responded with a wide range of answers. The most common specialties were pediatrics ($n = 7$) and ICU ($n = 6$), followed by family medicine and critical care ($n = 3$ each). Our subjects also included a neonatologist, a neurologist, an ICU cardiologist, a pediatric intensive care specialist, a neonatal nurse, two OB-GYNs, and several other miscellaneous specialties.

In addition to asking about their medical background, we asked subjects to report their level of expertise regarding traumatic brain injury (TBI) diagnosis and therapy. Subjects answered both questions on a five-point Likert scale, with “I know nothing about diagnosing/treating traumatic brain injury” and “I am an expert at diagnosing/treating traumatic brain injury” as the low and high anchor text. Figure 9.5 illustrates how subjects from different medical roles responded to these questions; unsurprisingly, attending physicians reported the highest levels of subject expertise.

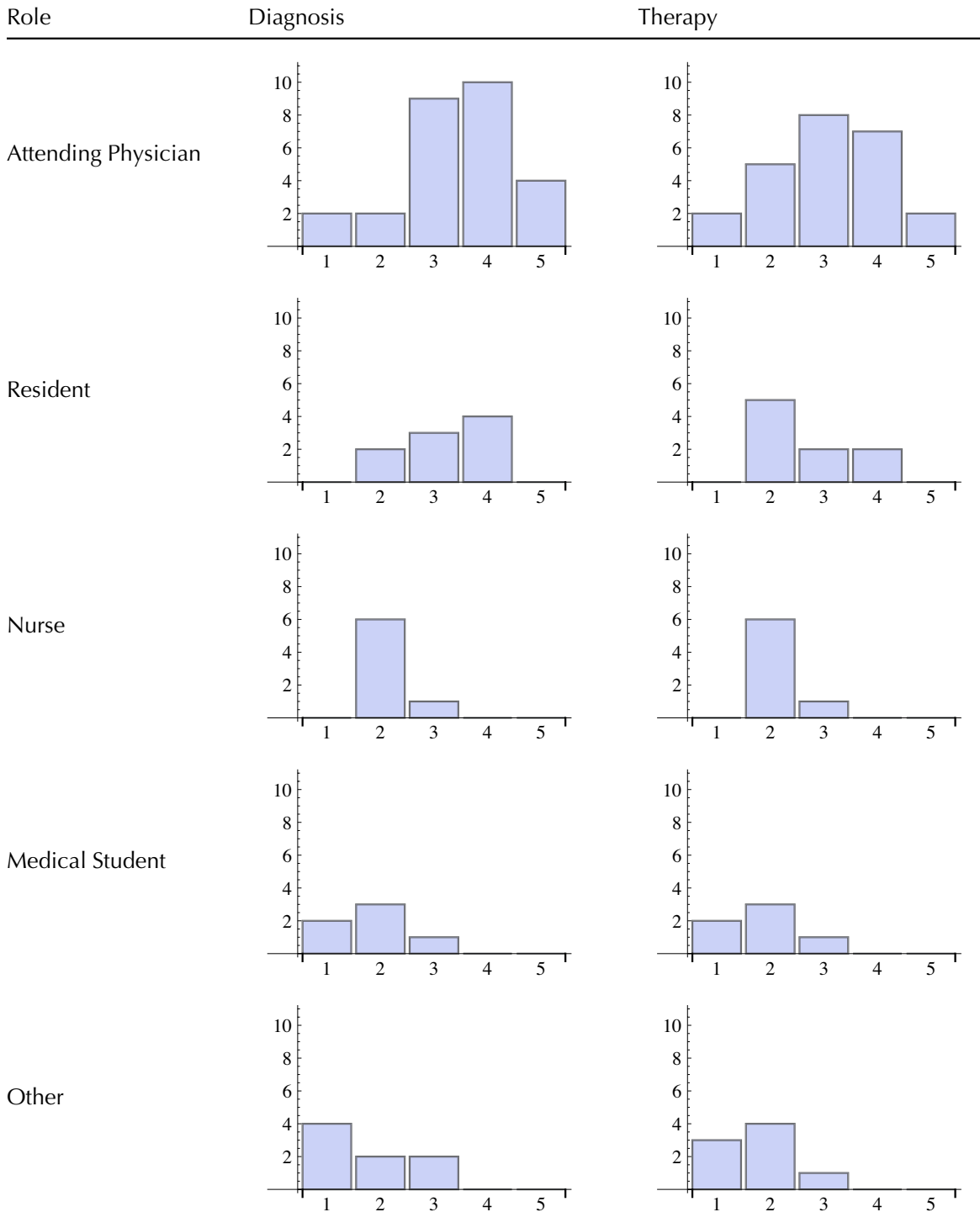


Figure 9.5: Reported expertise at diagnosing & treating traumatic brain injury, by medical role. Note the higher levels of expertise reported by attending physicians.

Role	<i>n</i>	%
Attending Physician	27	47.
Resident	9	16.
Nurse	7	12.
Medical Student	6	11.
Other	8	14.

Table 9.3: Subject count by medical role.

Role	Age	Std. Dev.
Attending Physician	41.2593	10.6142
Resident	29.7778	3.11359
Nurse	45.1429	7.71208
Medical Student	25.8333	4.70815
Other	41.25	9.00397

Table 9.4: Mean and standard deviations of subject ages, by medical role.

Computer & Search Background

Our questions about computer use and background were split into two sections: one about non-work-related computer use, and another about work-related computer use. Beginning with the non-work-related computer use questions, we see that our subjects were quite computer literate. Table 9.5 illustrates that 84% of our subjects reported using computers either daily or three–six times per week. Unsurprisingly, age seemed to play a role in computer use frequency (see Table 9.6), with younger subjects almost universally reporting daily non-work-related computer use.

In addition to being frequent computer users, our subjects were generally experienced at the use of Internet search engines. Virtually all subjects reported using search engines for non-work-related purposes at least once or twice per week (see Table 9.7), although, as with general computer use frequency, there was more variation among the older subjects than among the younger ones (see Table 9.8). Almost universally, subjects reported Google as their favorite search engine, with

only one subject reporting anything else (Yahoo).

When asked to rate, on a scale of one through five, their own ability to use Internet search engines to find non-work-related information³, subjects answered with an average response of 3.8—just above neutral, in other words (see Table 9.9 for the complete results). In contrast to the previous set of questions, older subjects tended to rate their searching ability more highly than younger subjects (see Table 9.10).

Approximately 79% of subjects reported using some kind of social networking software (e.g., Facebook, etc.); the majority of the subjects who did not were over 40 years of age, although it is important to note that far more older subjects used social networking systems than did not. Facebook was by far the most popular system, used by 91% of the subjects who reported using a social networking system; two subjects reported using Twitter, and one reported using MySpace.

Regarding work-related computer use, the picture is somewhat similar to that seen regarding non-work-related computer use. Roughly 72% of our subjects reported using a computer for work-related purposes on a daily basis, and a further 19% use one between three and six days per week (see Table 9.11); 79% of subjects reported using an Internet search engine to search for work-related information either 3–6 days per week, or daily (see Table 9.12). When asked to list the search engine they used most often for work-related searches, subjects exhibited a greater variety of responses than they did when asked about personal search engine use. While Google was still the most popular site (with 75% of subjects reporting it as their preferred search tool), nearly 10% of subjects reported preferring PubMed, and several other subjects mentioned both PubMed in addition to Google. Google Scholar was also a relatively popular choice.

Table 9.13 lists subjects' responses when asked to specifically identify online

³With "No ability" and "Expert searcher" as the anchor text.

Frequency	n	%
Every day	29	50.
Between 3 and 6 days per week	20	34.
1 or 2 days per week	7	12.
One or two days per month	2	3.4

Table 9.5: Non-work related computer use frequency.

resources they used specifically to look up medical information. Subjects were able to choose more than one answer. Electronic textbooks were popular, along with various incarnations of MEDLINE. The Cochrane Library was also popular, and subjects reported using both its English and Spanish versions. While many subjects marked “Other,” relatively few gave specifics. Of those that did, one mentioned UpToDate (but did not mention the language in which they typically used it), and another mentioned LILACS (a database of Spanish-language biomedical research literature from Latin America and the Caribbean.). When asked about the languages in which the electronic resources they used most were primarily published, 63% said “Mostly in English,” 27% said “Mostly in Spanish,” and the remainder said that they were approximately equally split between English and Spanish.

About 90% of our subjects reported having used an electronic dictionary or online translation tool such as Google Translate. Of the subjects who reported using such tools, 61% said that they used them on at least a weekly basis, with 19% reporting daily use (see Table 9.14).

Language Demographics

Almost without exception, our subjects reported regularly using only one language outside of work (97%), and that language was almost always Spanish. One subject reported using primarily Portuguese, another reported using both English and Spanish regularly, and a third reported using all three languages. One subject

	10-19	20-29	30-39	40-49
Every day	1	6	11	11
Between 3 and 6 days per week	0	2	8	10
1 or 2 days per week	0	1	4	2
One or two days per month	0	0	1	1

Table 9.6: Non-work-related computer use frequency, by age group. Note that virtually all of the subjects under the age of 30 reported daily use of computers, while older subjects exhibited more variation in their responses.

Frequency	n	%
Every day	24	41.
Between 3 and 6 days per week	20	34.
1 or 2 days per week	12	21.
One or two days per month	2	3.4

Table 9.7: Self-reported frequency of non-work-related Internet search engine use.

	10-19	20-29	30-39	40-49
Every day	1	3	10	10
Between 3 and 6 days per week	0	4	9	7
1 or 2 days per week	0	2	3	7
One or two days per month	0	0	2	0

Table 9.8: Non-work-related search engine use, by age. Note the oldest age group's amount of variation compared with that of the younger groups.

Response	n	%
1 - No Ability	0	0.
2	6	11.
3	24	42.
4	20	35.
5 - Expert Searcher	7	12.

Table 9.9: Subjects' self-reported ability to use Internet search engines to retrieve non-work-related information.

	10-19	20-29	30-39	40-49
1 - No Ability	0	0	0	0
2	0	1	2	3
3	0	4	11	9
4	1	2	6	11
5	0	1	5	1

Table 9.10: Subjects' search ability, by age. Note that older subjects tended to rate their search ability slightly more highly than younger subjects.

Frequency	n	%
Every day	42	72.
Between 3 and 6 days per week	11	19.
1 or 2 days per week	4	6.9
One or two days per month	1	1.7

Table 9.11: Work-related computer use frequency.

Frequency	n	%
Every day	26	45.
Between 3 and 6 days per week	20	34.
1 or 2 days per week	7	12.
One or two days per month	4	6.9
I don't use search engines for work-related Internet searches.	1	1.7

Table 9.12: Frequency with which subjects used Internet search engines for work-related purposes.

Resource	n
Electronic Textbook	45
MEDLINE (via PubMed)	43
Other	22
MEDLINE (via some other source)	13
Cochrane Library, Spanish	13
Cochrane Library, English	13
Other US CDC Reference	9
WHO-HINARI	5
Embase	3

Table 9.13: Electronic resources named by subjects when asked to select the resources they made use of for work-related purposes.

Frequency	n	%
Every day or almost every day	10	19.
Once or twice a week	22	42.
Once or twice per month	7	13.
Less often, but every once in a while	13	25.

Table 9.14: Frequency with which subjects who reported using online dictionary or translation tools reported doing so.

reported that they primarily used Italian. At work, however, the story was somewhat different. 14% of our subjects reported using two languages at work, with the second language always being English.⁴ Subjects reported a mean of five years of formal English education (with a standard deviation of 4.3 years).

9.2 Language Assessment Results

Recall from Section 8.1.2 that our primary language assessment outcome of interest was an estimate of the subjects' English reading proficiency, as determined by the Interagency Language Roundtable's reading self-assessment instrument. This instrument scores subjects on a scale of one through five, representing the five ILR reading proficiency "bands," which are officially designated as "R-0" through "R-4." The instrument also can potentially assign subjects a "plus" rating, indicating that they are somewhere in between two bands (i.e., a score of "3+" would indicate that the subject was somewhere between bands three and four).

For simplicity's sake, we did not generally make use of the "plus" score designations, and referred to each proficiency band as "1"–"5" (instead of the full, official designation of "R-0," etc.). Figure 9.7 shows the ILR score distribution (not counting "plus" levels) among our subjects as calculated by the routine shown in

⁴ When we repeated this particular analysis and included subjects who did not complete the study, we found three subjects who reported using Quechua (an indigenous language that is widely-spoken in parts of Peru and Bolivia) at work.

Appendix F, and Table 9.15 shows the actual numbers. The median band score was three, and this was also the most common score, into which 38% of subjects placed.

Some subjects skipped (i.e., did not indicate whether they agreed or disagreed) with one or more of the statements in the ILR instrument. Of our 59 subjects, nine skipped one or more of the questions in the ILR instrument (six skipped one question, and three skipped more than one). Initially, we were concerned that this would mean that we would be unable to use these subjects' responses to the instrument, and would be therefore unable to include them in our analysis, as missed questions would artificially lower these subjects' ILR scores (since questions with no answer do not count— or, rather, are counted in the same manner as negative responses— according to the ILR instrument's scoring rubric).

However, we realized that, due to the way in which the instrument's scoring works, it is entirely possible for a subject to skip questions in such a way that, no matter what their answer *would* have been had they answered the question, their score would have remained the same. Consider a subject who, due to their answers early in the instrument, was certain to test into one of the lower bands: if, for example, they agreed completely with the first five statements in the instrument and then began disagreeing with most of the rest of the statements. Now, imagine that they then skipped one of the instrument's final questions. Since their score is determined by the highest band with which they agree completely, any answer they could have given in the skipped question would not have affected their score. Figure 9.6 on page 182 illustrates this phenomenon.

In the figure, green squares represent statements in the ILR instrument that the subject agreed with, red squares represent statements that they disagreed with, and black squares represent statements to which they did not respond. The numbers represent the cutoffs between bands; questions 1 and 2 comprise the first band,

3–5 the second, and so on. Consider the case of subject 143, who skipped question number 14 as well as two of the final questions. Even if the subject agreed with all of these statements (instead of skipping them altogether), their score would have been unchanged, as they still would have disagreed with most of questions 11–14, which were what ultimately determined their score.

In the end, we determined that six of these subjects followed the pattern set by subject 143, and included their scores (and therefore the subjects themselves) in our analyses, since their scores remained unaffected by re-coding the missing questions as either “agree” or “disagree.” Three of the subjects, however, remained unusable (subjects 102, 132, and 216) as their scores changed depending on how their missing responses were recoded. Note that in each of these three cases, the issue was that one of the skipped questions happened to be at the boundary from one band to another; if these three subjects had agreed with those statements, their scores would have gone up from 2+ to 3 (or, in the case of subject 132, from 2+ to 3+).

In addition to the ILR instrument, we also assessed our subjects’ reading ability using a much simpler self-assessment: simply asking them to rate their ability to read English-language text (in both medical and non-medical contexts) on a scale of one through five, where one was “Cannot read any English” and five was “Perfect English reading ability.” The intention behind this question to compare subjects’ responses to this question to their responses to the more structured ILR instrument. Figure 9.8 shows the distribution of responses for reading in non-medical contexts, and Figure 9.9 shows the distribution for medical contexts. Subjects rated their ability to read English in a medical context higher than they rated their ability to read English in non-English contexts (Sign-Rank test, $p < 0.001$).

Subjects’ ILR scores were significantly correlated with their responses to the simple self-report questions ($r^2 = 0.61$, $p < 0.001$, same for both medical and non-

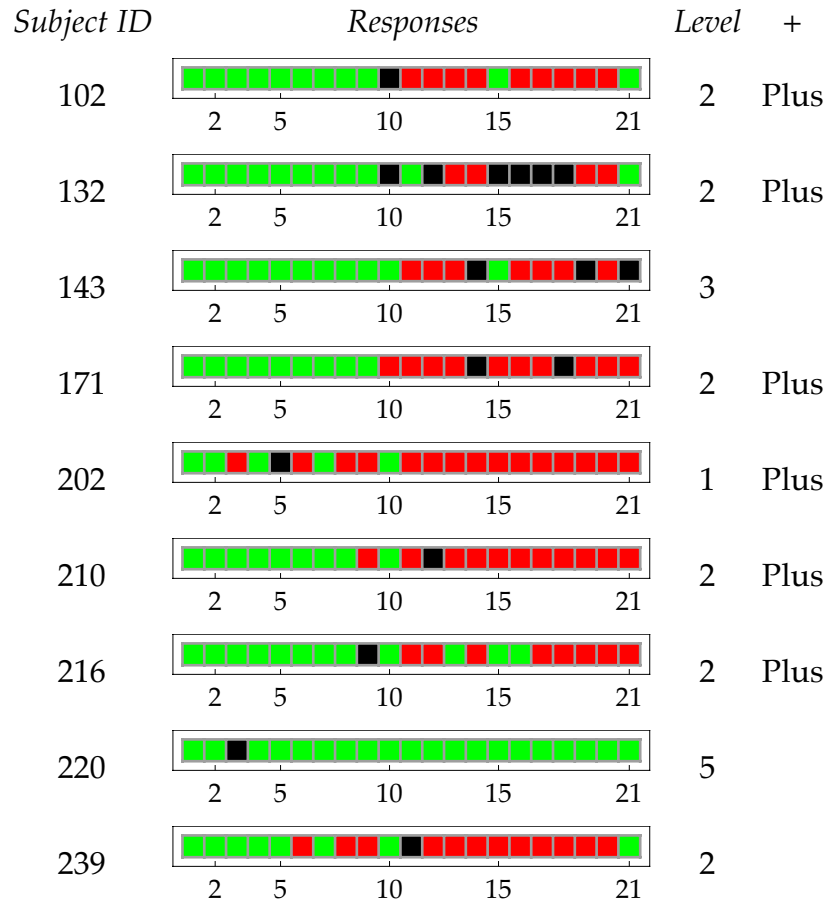


Figure 9.6: The responses of the nine subjects who skipped at least one of the ILR instrument's questions, along with their scores. Green squares represent statements in the ILR instrument that the subject agreed with, red squares represent statements that they disagreed with, and black squares represent statements to which they did not respond. The numbers indicate where the instrument's rating cutoffs are— for example, questions one and two comprise the first band's statements, while questions three through five (inclusive) represent the second band's, and so on. We were ultimately able to include six of these subjects' scores in our analyses, since their scores remained unaffected by re-coding the missing questions as either "agree" or "disagree." Three of the subjects, however, remained unusable (subjects 102, 132, and 216) as their scores changed depending on how their missing responses were recoded.

Level	n	%
1	6	11.
2	7	12.
3	21	38.
4	13	23.
5	9	16.

Table 9.15: ILR Score distribution among subjects.

Rating	n	%
Lower than average	19	32.
About average	18	31.
Higher than average	22	37.

Table 9.16: Subjects assessment of how their own English reading abilities compare to those of their colleagues.

medical contexts). ILR scores were also weakly (but significantly) correlated with subjects' English educational background ($r^2 = 0.28$, $p < 0.001$, see Figure 9.10).

When asked to indicate how they rated their own English reading abilities as compared to those of their colleagues, subjects were nearly evenly split between those who said their English abilities were "Lower than average" among their colleagues, "About average," and "Higher than average" (see Table 9.16). Stratifying responses to this question by ILR score, however, is instructive (see Figure 9.11). The figure shows that subjects in higher ILR bands were more likely to rate their own abilities as being "above average," while subjects lower on the ILR scale were more likely to rate themselves as being "below average."

ILR did not vary significantly by role (Kruskal-Wallis test, $p = 0.15$), although nurses did have a lower median ILR score than the other roles (2 vs 3, see Table 9.17). However, higher ILR level was associated with higher self-reported levels of computer expertise (Fisher's exact test, $p = 0.004$).

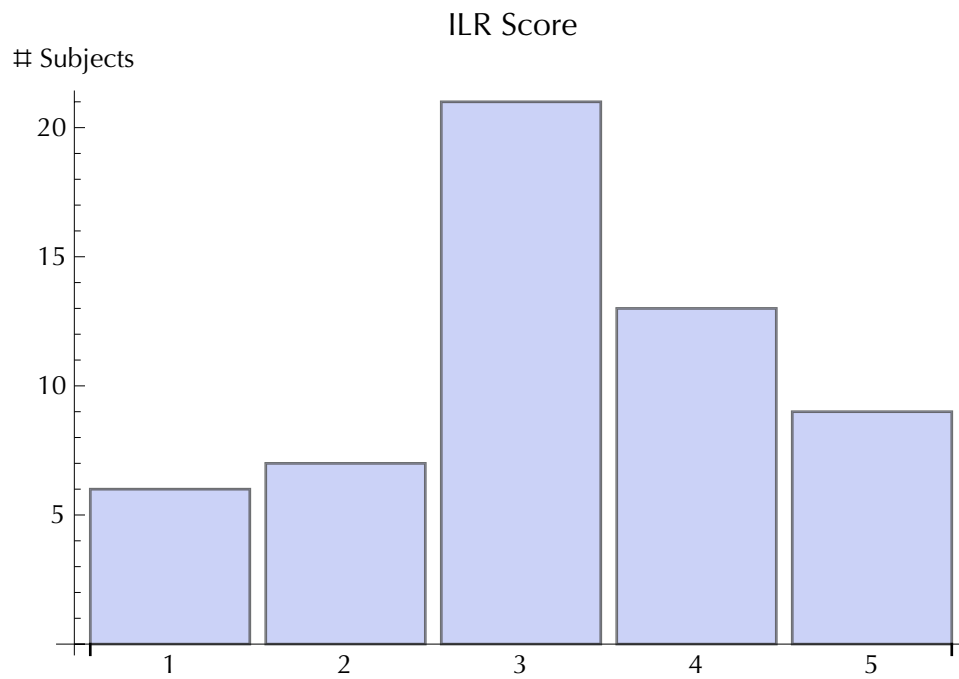


Figure 9.7: Distribution of ILR score among subjects.

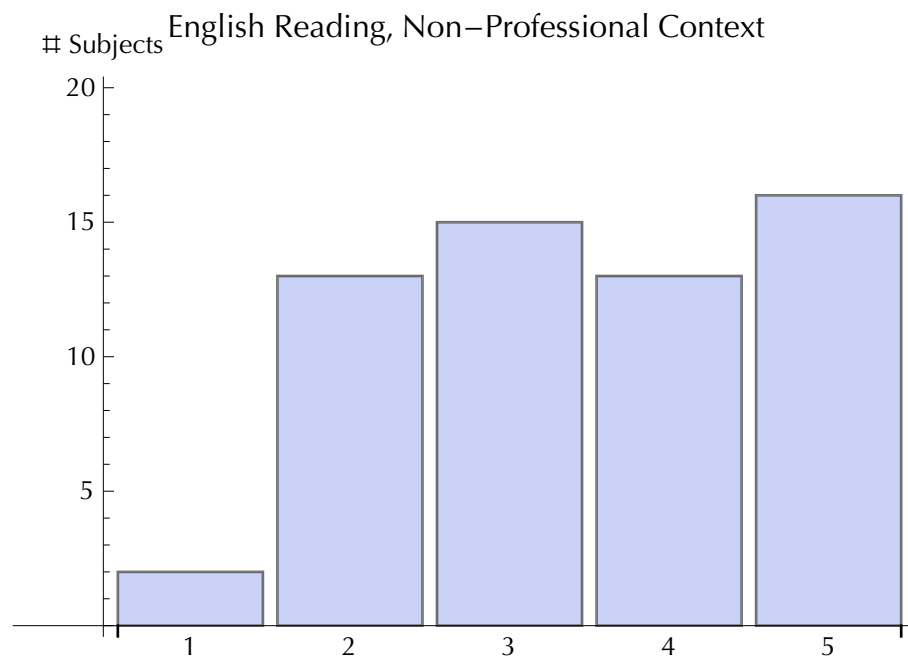


Figure 9.8: Distribution of self-assessed proficiency for reading English in non-medical contexts.

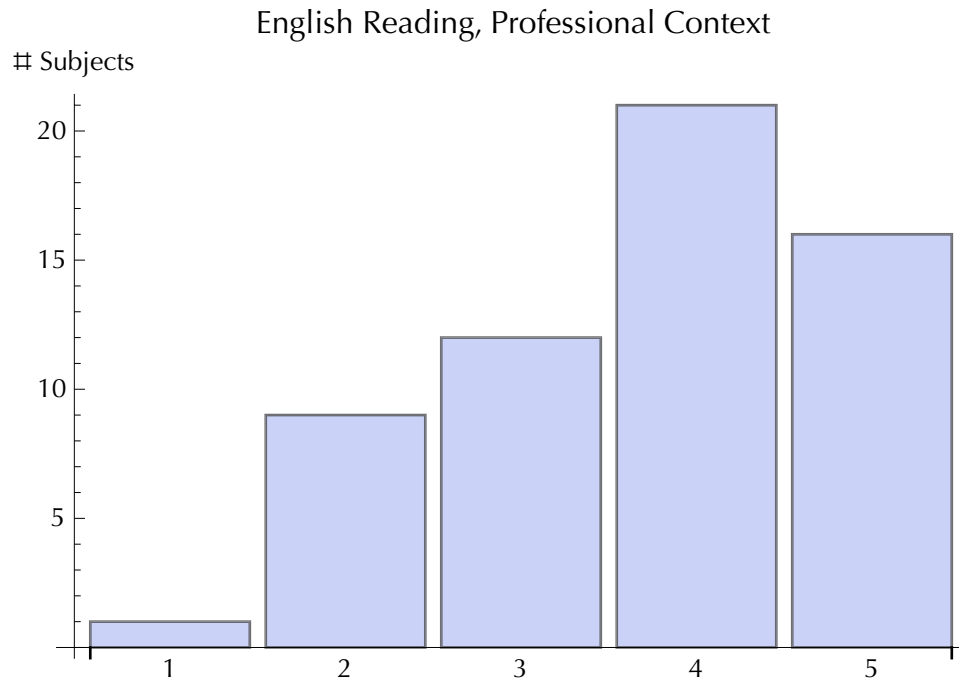


Figure 9.9: Distribution of self-assessed reading proficiency for for reading English in medical contexts.

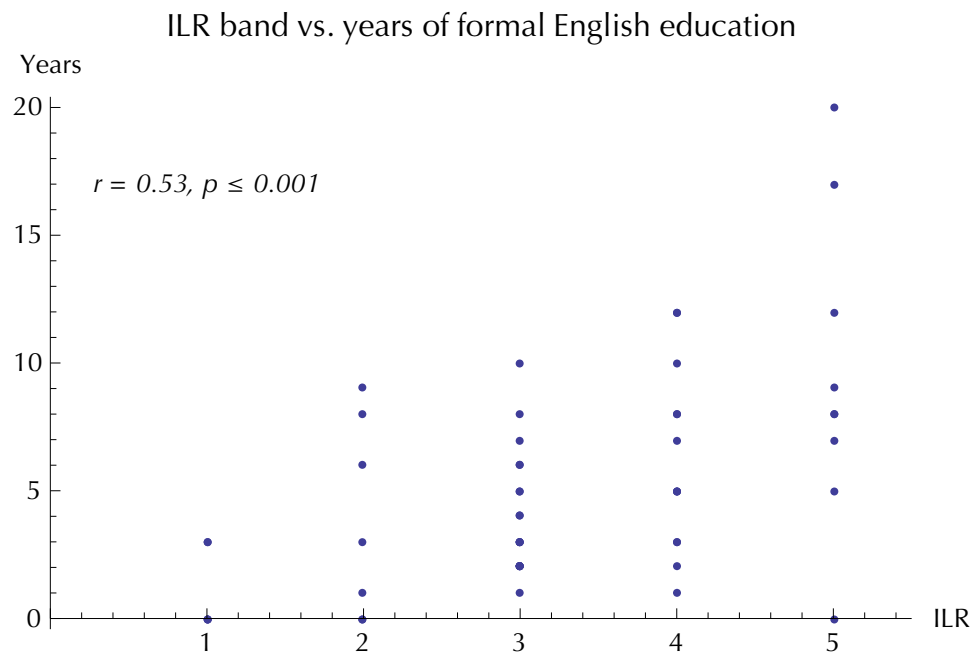


Figure 9.10: Subject ILR scores vs. years of formal English education. Note the weak but significant correlation.

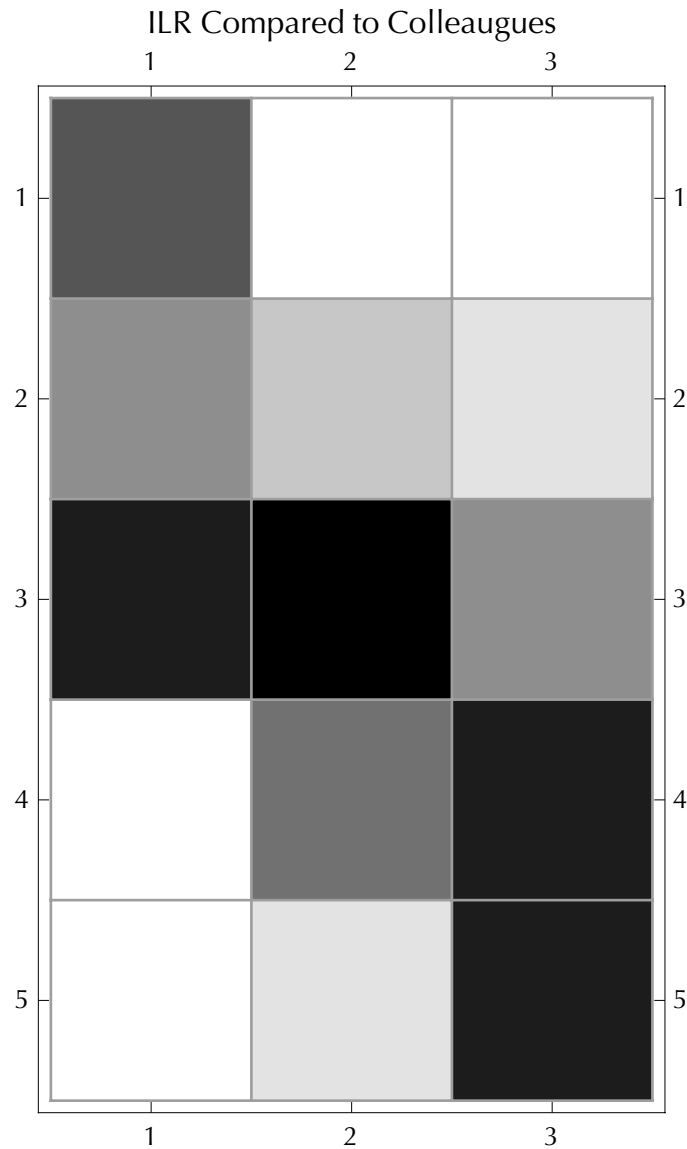


Figure 9.11: Data from Table 9.16, stratified by ILR score. Rows represent ILR scale levels, and columns represent responses to the colleague-comparison question; a cell's shading represents the number of subjects in that particular ILR band who replied to the comparison question in that particular way. The first column represents subjects who responded with "Lower than average," the middle column "About average," and the third column "Higher than average." Note that subjects who scored lower on the ILR scale tended to rate themselves as "Lower than average," while subjects higher on the ILR scale tended to rate themselves more highly.

Role	Mean ILR	Median ILR
Attending Physician	3.35	3
Resident	3.44	3
Nurse	2.17	2
Medical Student	3.4	3
Other	3.25	3

Table 9.17: Medical role vs. mean and median ILR scores. While there was no significant difference across roles (Kruskal-Wallis $p = 0.15$), it is worth noting that nurses had lower median ILR scores than subjects in the other categories.

9.3 Subject Preferences

Recall from Section 8.3 that our main metrics for subject preferences regarding interface mode were the subjects' responses to a series of questions indicating which of the three interface modes ("English-only," "Spanish-only," and "Bilingual") they found to be easiest, most efficient, etc. Figure 9.12 shows the subjects' unstratified responses. Each row represents one of the usability feedback questions (described in Section 8.1.3; the complete text of the English-language version of the instrument may be found Appendix I); each region indicates the percentage of subjects who replied that the a particular interface was easiest/fastest/etc., with blue, green, and orange regions representing the English-only, Bilingual, and Spanish-only interfaces, respectively.⁵ Note that the first four horizontal bars represent "positively"-worded questions (e.g., 42.4% of subjects liked the English-only interface the best) while the fifth bar represents a "negatively"-worded question (51% of subjects found the English-only interface to be the most difficult one to use).

The first thing to notice regarding this figure is that the regions are not equally sized; that is to say, subjects developed and expressed discernible preferences between the interfaces. The bilingual interface was quite popular, but was often quite

⁵ For the benefit of readers who may not be able to see the colors, the regions are described in the same order that they appear in the figure: blue/English-only, green/Bilingual, and orange/Spanish-only, from left to right.

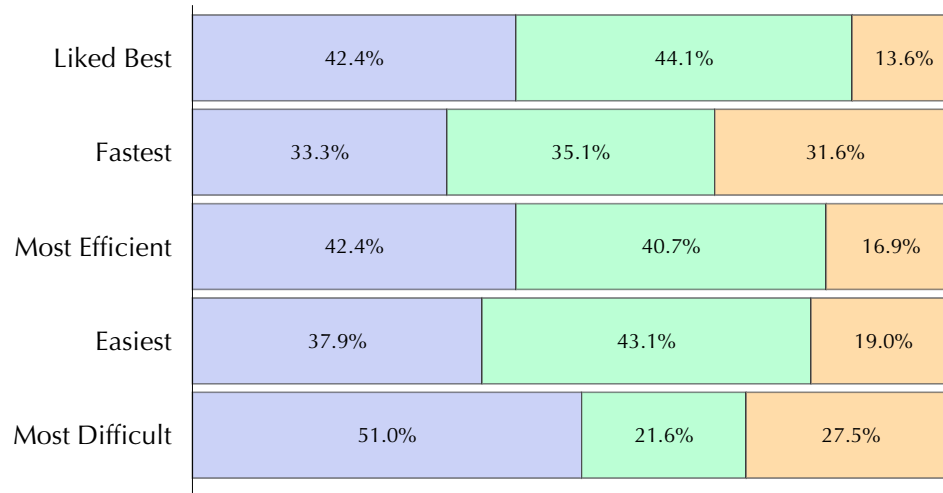


Figure 9.12: Subjects' unstratified responses to the usability feedback questions, indicating the percentages of subjects who answered that the English-only (blue), Bilingual (green), and Spanish-only (orange) system was easiest/fastest/etc.

close in popularity to the English-only interface. Subjects were most likely to think of the English-only interface as being the most difficult one to use.

Our subjects' responses to the positively-worded questions were relatively self-consistent. In other words, subjects who reported finding the English-only system to be "easiest" were most likely to report that they "liked" that system best, and so on. Tables 9.20 through 9.23 (pages 194 through 197) report the raw cross-tabulated data; note the strong degree of diagonal homogeneity running through the tables representing the "positively-worded" questions. Note further the heterogeneity in subjects' responses to the "negatively-worded" question— subjects rarely felt that the most "difficult" system was also the easiest one to use, etc.

Another interesting piece of heterogeneity took the form of "cross-talk" between the subjects' feelings regarding the Bilingual and Spanish-only interfaces. For example, there was some notable variation among the 18 subjects who felt *fastest* at using the Spanish-only system in terms of which system they found to be easiest to use: while nine of these subjects found the Spanish-only interface to be easiest to use, 7 found the bilingual interface to be easiest (Table 9.20.B). In

the same table, the subjects who reported finding the Bilingual interface *easiest* to use were split 16/7 in terms of which interface (between Bilingual and Spanish-only) they felt was the *fastest* one to use. This stands in contrast to the relatively self-consistent behavior of the 19 subjects who felt *fastest* at using the Bilingual interface, 16 of whom found the Bilingual interface to be *easiest* to use.

This behavior exhibited itself again most strongly in the Fastest/Efficient crosstabulation (see Table 9.22). Apparently, subjects who felt fastest at using the Bilingual interface were generally consistent in their opinions of which interface was easiest/efficient/etc.—but not necessarily vice versa.

In addition to examining at the direct preference data for each interface mode, we grouped the “Spanish-only” and “Bilingual” results together, and prepared a “Some Spanish”/“No Spanish” contrast. The contrasted results for the feedback questions can be seen in Table 9.18; note that while the Spanish-containing interfaces tended to be more popular than than the English-only interface, the only question for which the Some Spanish/No Spanish split was significantly different from 50/50 was that of which mode was fastest (two-sided $p = 0.015$, Table 9.18.C).⁶

Subject preferences were heavily affected by their English proficiency. In short, subjects with stronger English reading abilities (as measured by the ILR instrument) were much more likely to prefer the English-only interface than either of the other interfaces, and this likelihood increased with ILR score. Table 9.19 and Figure 9.13 (pages 191 and 198, respectively) gives the numerical results and illustrate this trend graphically. Each plot’s rows represent ILR levels, and the columns represent the three possible responses to the various usability feedback questions (English-only, Bilingual, and Spanish-only). A given cell’s shading indicates the number of subjects in that particular ILR level who felt that particular interface

⁶ Although the question of which interface mode was easiest was perhaps marginally significantly different (two-sided $p = 0.09$, table Table 9.18.A).

<i>A. Easiest</i>			<i>B. Efficient</i>		
Mode	<i>n</i>	<i>%</i>	Mode	<i>n</i>	<i>%</i>
Some Spanish	36	62.	Some Spanish	34	58.
English-only	22	38.	English-only	25	42.

<i>C. Fastest</i>			<i>D. Liked</i>		
Mode	<i>n</i>	<i>%</i>	Mode	<i>n</i>	<i>%</i>
Some Spanish	38	67.	Some Spanish	34	58.
English-only	19	33.	English-only	25	42.

<i>E. Difficult</i>		
Mode	<i>n</i>	<i>%</i>
Some Spanish	25	49.
English-only	26	51.

Table 9.18: Dichotomized responses to the usability feedback questions.

mode, with darker cells containing higher numbers of responses. Thus, in Figure 9.13.A, we see that no subjects in ILR bands 1 and 2 thought that the English-only system was easiest, and likewise no subjects in bands 4 and 5 thought that the Spanish-only system was easiest.

The effect of English proficiency on user preferences is even more pronounced when we examine the Some-Spanish/English-only contrast described above. Figure 9.14 illustrates this finding; note that, for example, the proportion of subjects who felt that either the Spanish-only or Bilingual interfaces were easiest is much higher among subjects in ILR bands 1 and 2, and falls steadily in the higher ILR bands. Note, however, that the proportion does not reach zero, implying that, even at the highest levels of English reading proficiency, many subjects still felt that one of the two Spanish-containing interfaces was easiest to use.

When we examine a “Spanish-only”/“Other” contrast, we see a similar pattern. Subjects with weaker English reading proficiency were more likely to prefer

A. Easiest				B. Efficient			
	English-only	Bilingual	Spanish-only		English-only	Bilingual	Spanish-only
ILR 1	0	3	3	ILR 1	0	3	3
ILR 2	0	5	2	ILR 2	2	3	2
ILR 3	7	8	5	ILR 3	9	8	4
ILR 4	9	4	0	ILR 4	8	5	0
ILR 5	5	4	0	ILR 5	5	4	0

C. Fastest				D. Liked			
	English-only	Bilingual	Spanish-only		English-only	Bilingual	Spanish-only
ILR 1	0	2	4	ILR 1	0	4	2
ILR 2	0	5	2	ILR 2	0	6	1
ILR 3	6	6	8	ILR 3	11	7	3
ILR 4	7	3	2	ILR 4	8	4	1
ILR 5	5	3	1	ILR 5	5	4	0

E. Difficult			
	English-only	Bilingual	Spanish-only
ILR 1	6	0	0
ILR 2	6	0	0
ILR 3	6	5	6
ILR 4	3	3	5
ILR 5	5	1	2

Table 9.19: Subject responses to the usability feedback questions, stratified by ILR level (English reading proficiency).

the Spanish-only interface to either of the other two interfaces, and this tendency decreased with increased ILR ability (see Figure 9.16). In fact, virtually no subjects in bands 4 and 5 (i.e., the subjects who were most proficient at English reading) expressed a preference for the Spanish-only interface. Figure 9.17 shows that our subject's likelihood of preferring the Bilingual interface, however, did not seem to depend on ILR score as strongly, although in general subjects with lower ILR did seem to prefer the Bilingual mode more often than subjects with stronger reading proficiencies.

These data illustrate several clear trends. First, many subjects preferred one of the interfaces containing translated content to the interface with English-only content. Second, this preference was related to subject English reading proficiency, particularly when considering the subjects who preferred the Spanish-only interface to the other two. Third, the Bilingual interface, while not clearly more popular than the other two, was still preferred by a sizable contingent of the subjects, and subjects' language-proficiency-based preferences for (or against) the Bilingual mode did not seem to follow the same pattern as did subjects' preference for the other two modes. That said, however, it is important to keep in mind that the sample sizes between bands are very unequal, and that relatively few subjects tested into the lowest English proficiency bands. As such, it is difficult to draw statistically meaningful conclusions from these data.

It should be noted here that, although many subjects expressed preferences for translated over untranslated content, virtually all of our subjects (including many of those who preferred one of the translated interfaces) voiced complaints about the *quality* of the translations. Some simply said that the translations could be difficult to follow, whereas others specifically mentioned that the translations contained outright errors. Many subjects even said that the errors impeded their ability to understand the contents of the articles. At the same time, however, sub-

jects were quite clear about the fact that the translated content had value. Several subjects reported being able to grasp the overall content of a search result when using one of the translated modes, and many mentioned that they found reading in their native language to be easier than reading in English.

		A. Efficient		
		English-only	Bilingual	Spanish-only
Easiest	English-only	20	2	0
	Bilingual	3	19	3
	Spanish-only	1	3	7

		B. Fastest		
		English-only	Bilingual	Spanish-only
Easiest	English-only	18	1	2
	Bilingual	1	16	7
	Spanish-only	0	2	9

		C. Liked		
		English-only	Bilingual	Spanish-only
Easiest	English-only	20	1	1
	Bilingual	2	21	2
	Spanish-only	2	4	5

		D. Difficult		
		English-only	Bilingual	Spanish-only
Easiest	English-only	4	5	11
	Bilingual	15	4	2
	Spanish-only	7	2	1

Table 9.20: Cross-tabulation of subject responses between the “Easiest” usability feedback question and the other four questions. Note that the cell values are counts of subjects who responded in a particular way; for example, 20 subjects found the English-only system to be both easiest and most efficient to use. Note the considerable diagonal homogeneity within the four “positively-worded” questions, and the heterogeneity in the responses to the “negatively-worded” question.

		A. Easiest		
		English-only	Bilingual	Spanish-only
Efficient	English-only	20	3	1
	Bilingual	2	19	3
	Spanish-only	0	3	7

		B. Fastest		
		English-only	Bilingual	Spanish-only
Efficient	English-only	16	4	4
	Bilingual	3	14	6
	Spanish-only	0	2	8

		C. Liked		
		English-only	Bilingual	Spanish-only
Efficient	English-only	22	2	1
	Bilingual	2	21	1
	Spanish-only	1	3	6

		D. Difficult		
		English-only	Bilingual	Spanish-only
Efficient	English-only	6	6	10
	Bilingual	13	4	3
	Spanish-only	7	1	1

Table 9.21: Cross-tabulation of subject responses between the “Efficient” usability feedback question and the other four questions. See the caption for Table 9.20 for further explanation.

		A. Easiest		
		English-only	Bilingual	Spanish-only
Fastest	English-only	18	1	0
	Bilingual	1	16	2
	Spanish-only	2	7	9

		B. Efficient		
		English-only	Bilingual	Spanish-only
Fastest	English-only	16	3	0
	Bilingual	4	14	2
	Spanish-only	4	6	8

		C. Liked		
		English-only	Bilingual	Spanish-only
Fastest	English-only	18	1	0
	Bilingual	2	18	0
	Spanish-only	4	6	8

		D. Difficult		
		English-only	Bilingual	Spanish-only
Fastest	English-only	4	4	9
	Bilingual	12	3	2
	Spanish-only	10	3	2

Table 9.22: Cross-tabulation of subject responses between the “Fastest” usability feedback question and the other four questions. See the caption for Table 9.20 for further explanation.

		A. Easiest		
		English-only	Bilingual	Spanish-only
Liked	English-only	20	2	2
	Bilingual	1	21	4
	Spanish-only	1	2	5

		B. Efficient		
		English-only	Bilingual	Spanish-only
Liked	English-only	22	2	1
	Bilingual	2	21	3
	Spanish-only	1	1	6

		C. Fastest		
		English-only	Bilingual	Spanish-only
Liked	English-only	18	2	4
	Bilingual	1	18	6
	Spanish-only	0	0	8

		D. Difficult		
		English-only	Bilingual	Spanish-only
Liked	English-only	4	6	11
	Bilingual	17	3	2
	Spanish-only	5	2	1

Table 9.23: Cross-tabulation of subject responses between the “Liked” usability feedback question and the other four questions. See the caption for Table 9.20 for further explanation.

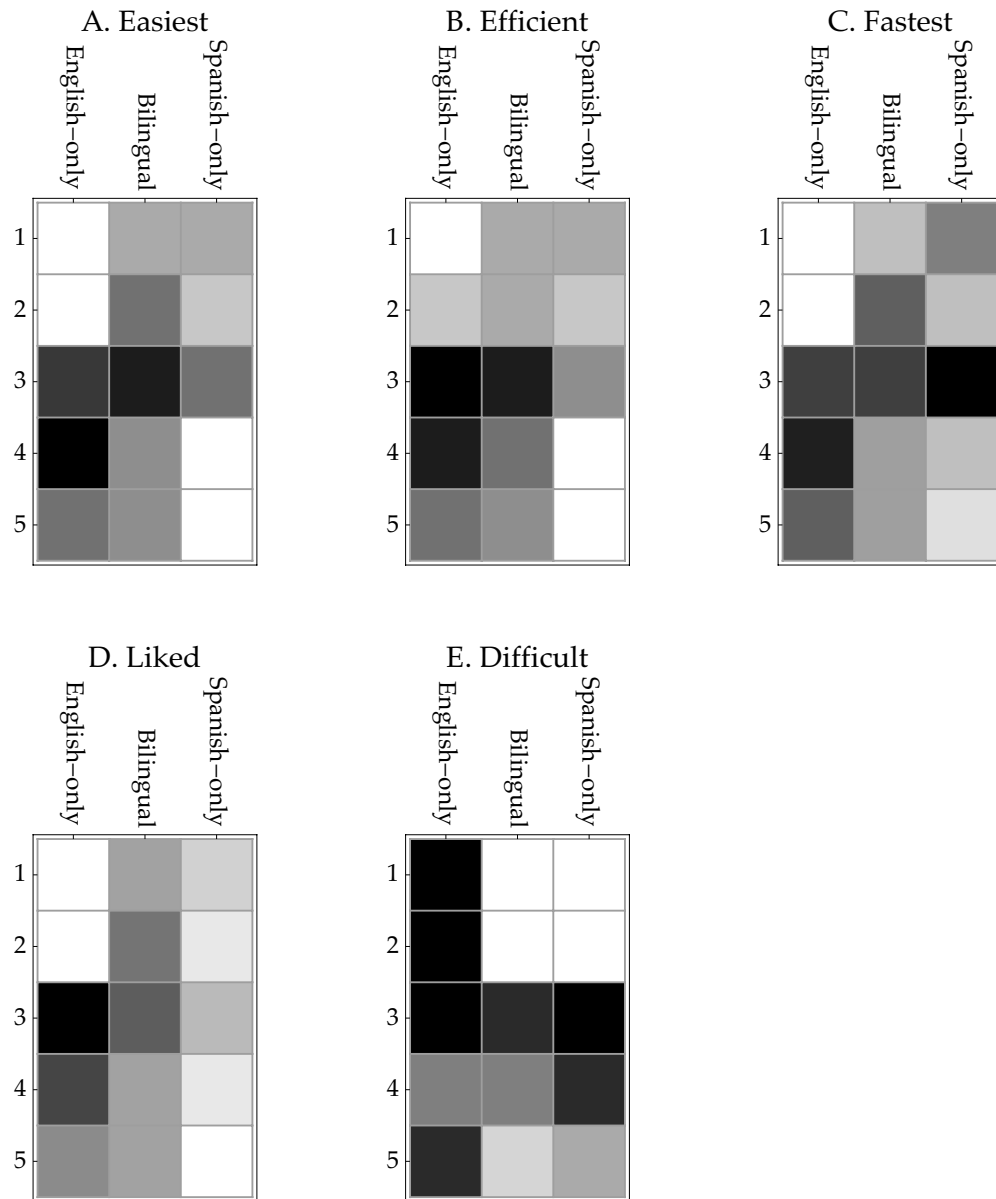


Figure 9.13: The same data as Table 9.19, rendered as density plots. Cell shading indicates the number of subjects in a given ILR level who felt that a given interface was easiest/fastest/etc.; the darker the cell, the higher the number of subjects. Note the strong diagonal symmetry in the plots representing the four positively-worded questions: very few (if any) subjects in the lower ILR bands found the English-only interface to be easiest, fastest, most efficient, or likable; likewise, the Spanish-only interface was quite unpopular with subjects with higher levels of English reading proficiency.

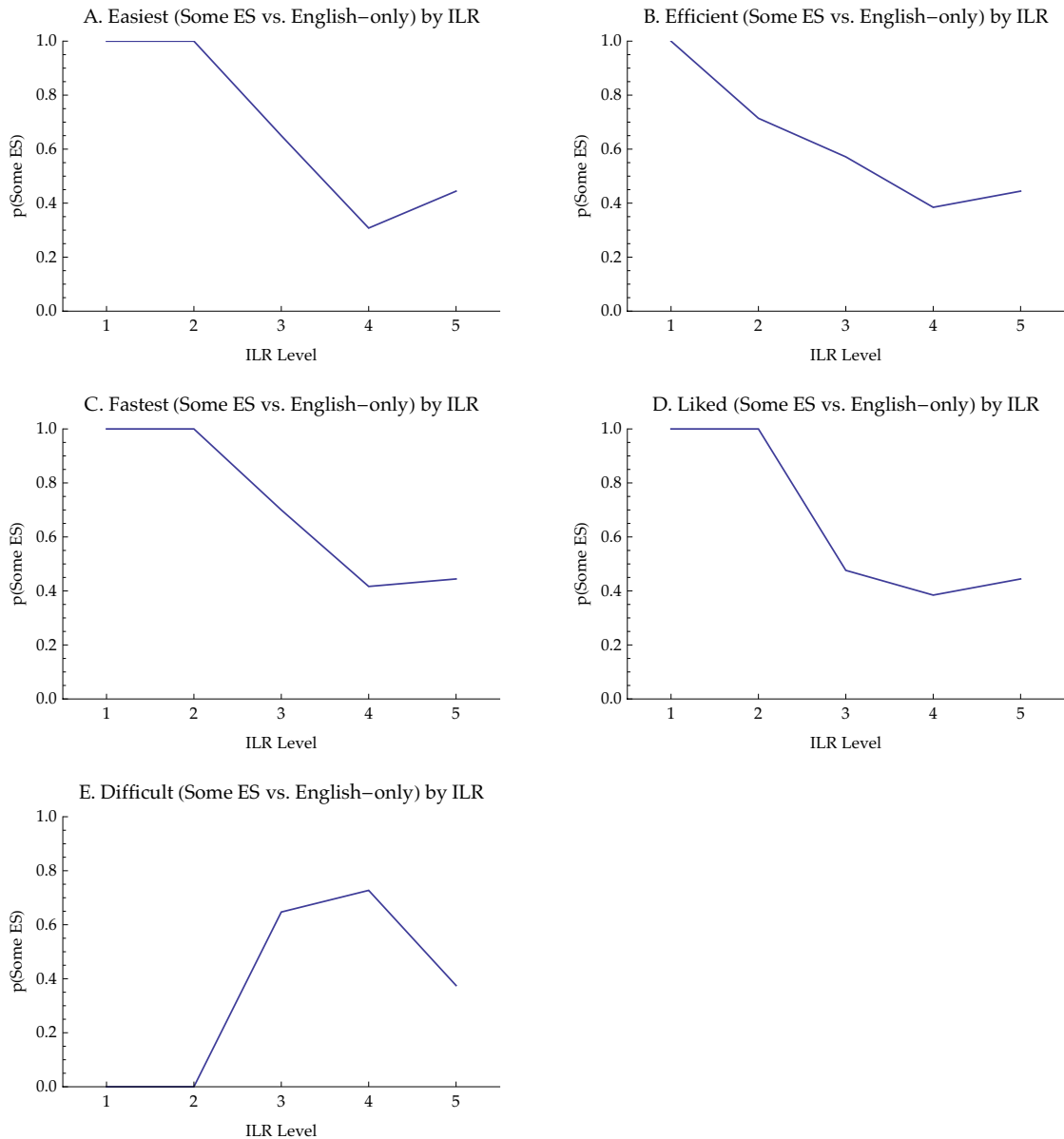


Figure 9.14: Subject responses to the usability feedback questions, dichotomized into “Spanish-only”/“Bilingual” (“Some Spanish”) vs. “English-only.” Abscissae are ILR levels; ordinates represent the proportion of subjects in that ILR level to respond with either “Spanish-only” or “Bilingual.” Note that, as ILR levels increase, the number of subjects preferring “Some-spanish” decreases.

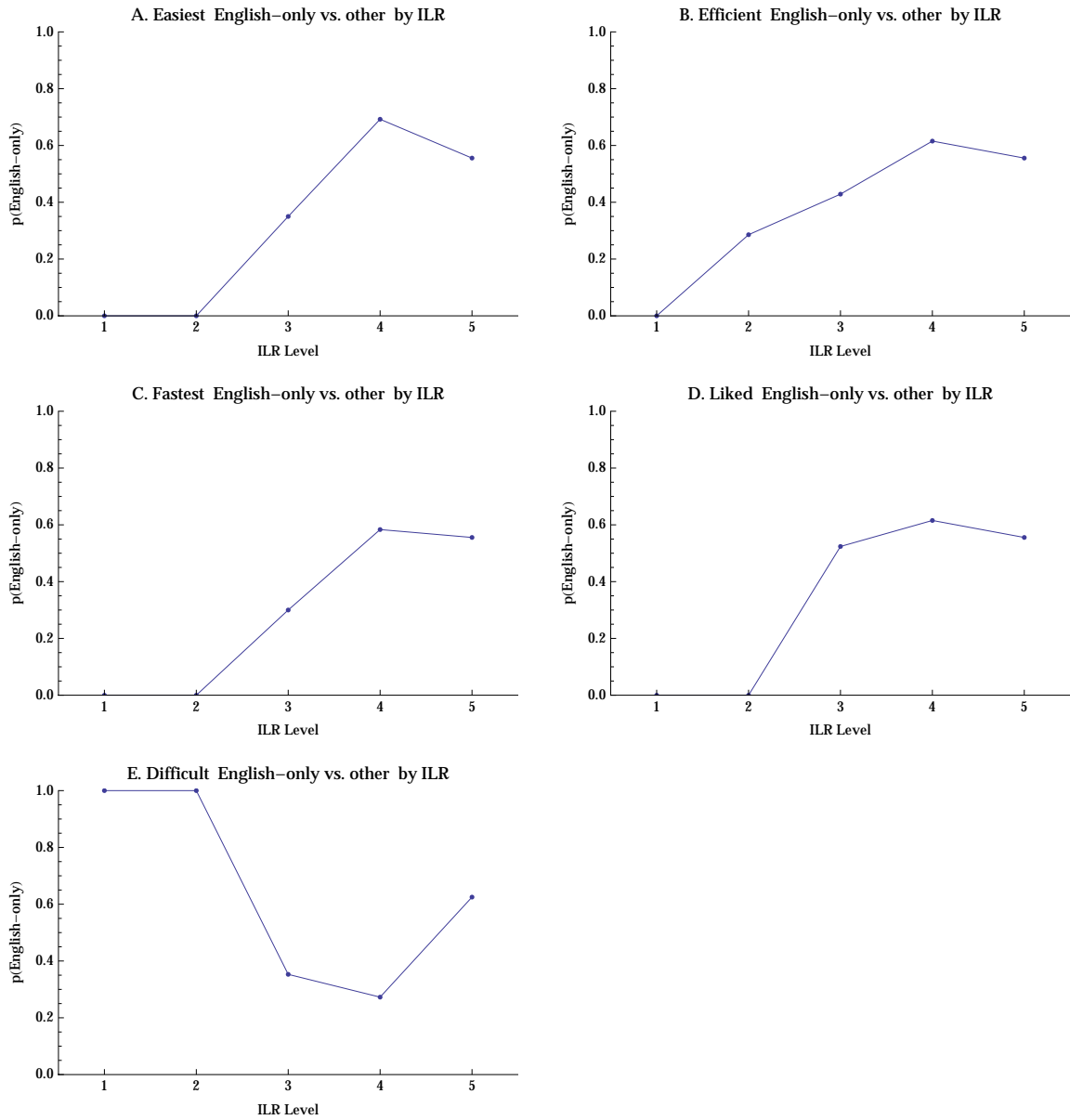


Figure 9.15: Proportion of subjects preferring the “English-only” interface to the other two choices for the five usability feedback questions, stratified by ILR. Note that the graphs in this figure are essentially inverted versions of those shown in Figure 9.14.

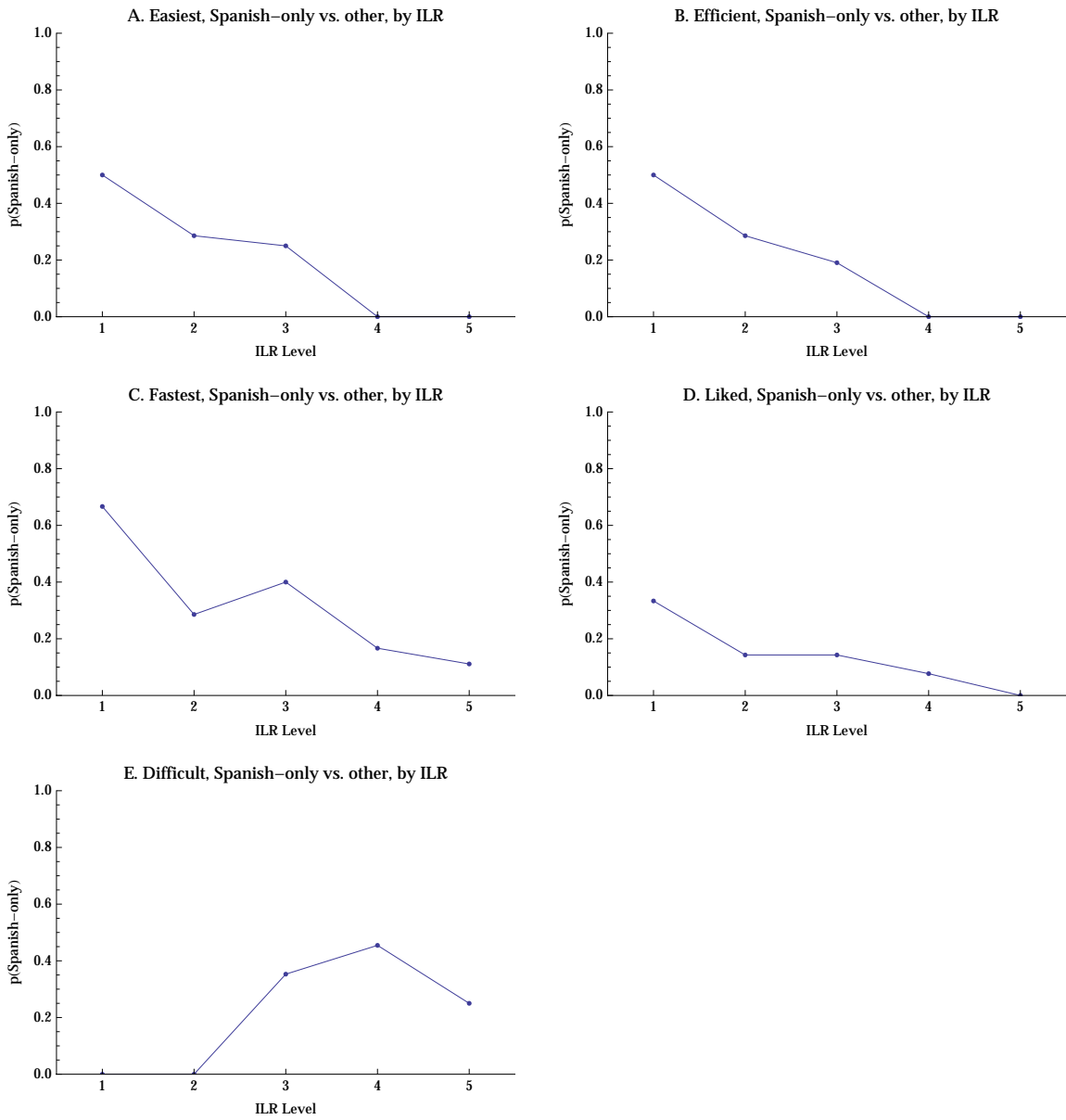


Figure 9.16: Proportion of subjects preferring the “Spanish-only” interface to the other two choices for the five usability feedback questions, stratified by ILR.

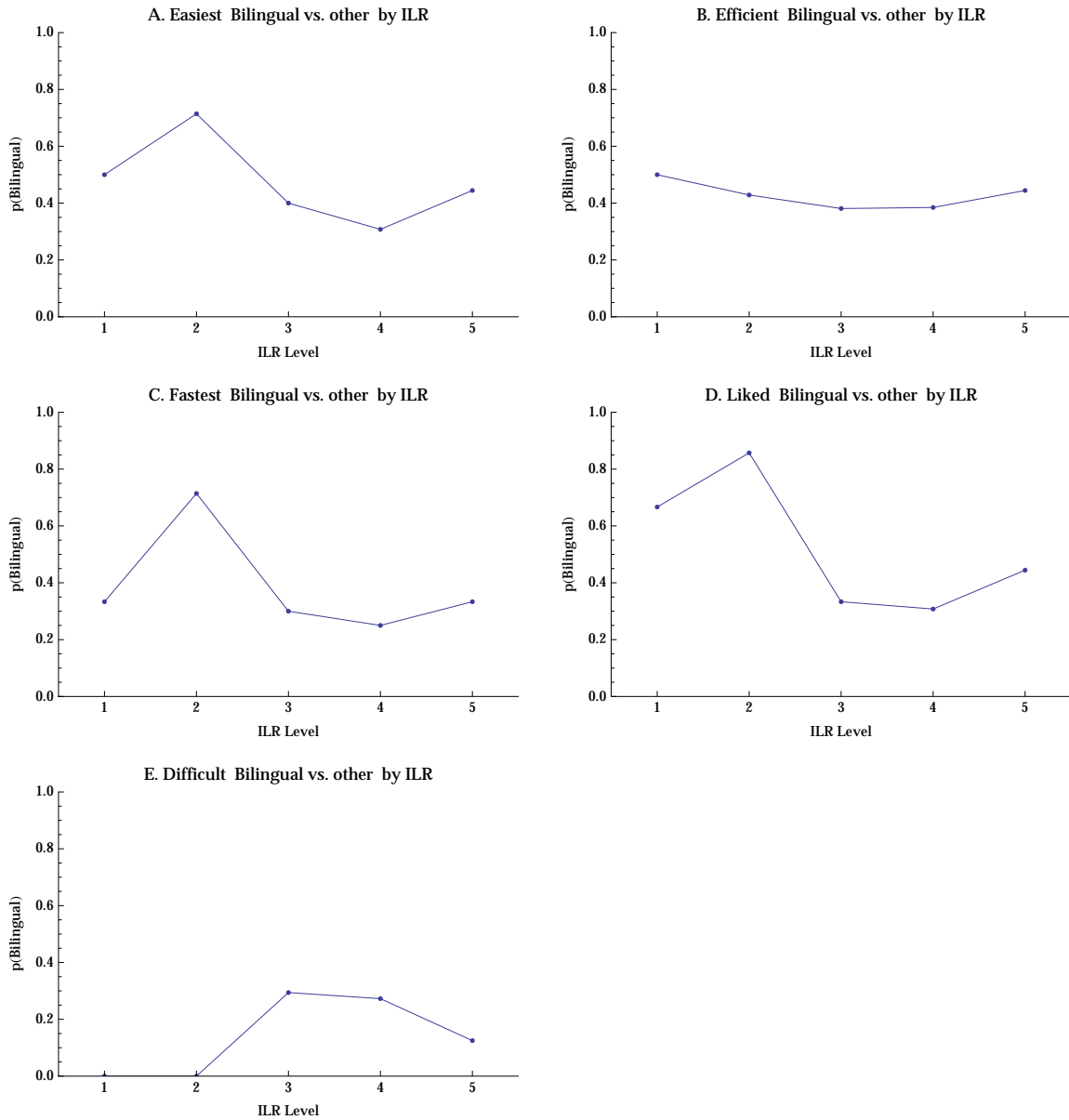


Figure 9.17: Proportion of subjects preferring the “Bilingual” interface to the other two choices for the five usability feedback questions, stratified by ILR. Note that subjects’ opinions regarding this interface mode were more variable than their opinions of either of the two monolingual modes, and that these figures do not exhibit the same ILR-level-related trends as were seen in figures 9.15 and 9.16.

9.4 Subject Performance

As described in Section 8.3, we considered two dimensions of subject performance: the accuracy with which they were able to identify documents about traumatic brain injury, and the speed at which they were able to do so. This section describes the results for each dimension in turn.

9.4.1 Classification Accuracy

Recall that the purpose of the document selection task was to simulate a realistic use scenario (in this case, reviewing the results retrieved by an Internet search engine, and deciding which results were relevant to a specific information need) under several different interface conditions, specifically, a Monolingual English condition (Mono EN), a Monolingual Spanish (Mono ES) condition, and a Bilingual (Bi) condition. The subjects reviewed 40 articles per condition, approximately 25% of which were “relevant.”⁷ Subjects exhibited a wide amount of variation in terms of how many articles they selected (i.e., some subjects selected many articles, other selected fewer, and so on); see Figure 9.18 on page 207.

That said, there was strong within-subject consistency— subjects who selected many articles under one interface tended to select relatively many under the other interfaces (see Figure 9.19). Selection counts differed by interface mode (Friedman test, $p < 0.001$); Subjects tended to mark more articles as relevant (“select more articles”) when using the Spanish-language interface than when using the English-only interface (Wilcoxon Sign-Rank test, $p < 0.001$). Table 9.24 (page 208) lists the mean, median, and standard deviation of the selection counts for each topic.

Using one-way ANOVA, we determined that subject expertise in both TBI di-

⁷ Due to how we generated the document collection, there was unexpected variability between document piles in terms of the precise number of documents that were from the “relevant” and “not-relevant” sets, although the numbers were approximately equal.

agnosis nor TBI therapy expertise (as measured by subjects' response to questions about their level of knowledge regarding traumatic brain injury therapy and diagnosis; see Section 8.1.1) was unrelated to the number of articles selected by subjects; similar analysis revealed that ILR was similarly unrelated to the number of articles selected by subjects. Two-way ANOVA using various combinations of TBI diagnosis, therapy, and ILR level also failed to find any association or interaction between any of those factors and subject selection counts.

As previously discussed, our metrics for classification accuracy were precision, recall, and F-measure (see Section 8.3). There was considerable between-subject variation in each of those metrics; Figures 9.20 through 9.22 show the distribution of the three metrics across each task. Since the subjects' performance measures were not normally distributed, we primarily used non-parametric methods where appropriate for our analysis, and in this discussion we generally refer to median values instead of mean values. Table 9.25 presents the median precision, recall, and F-measures for each task, together with standard deviations.⁸ Note the generally low precisions and F-measures, as well as the high standard deviations.

Across all three measures, subjects seem to have performed best under the Monolingual English interface, followed by the Bilingual and Monolingual Spanish interfaces; the Wilcoxon Signed-rank test shows that subjects' F-measures were higher under the English interface than they were under either the Monolingual Spanish or Bilingual interfaces (one-sided $p < 0.001$ for both), and further indicates that subjects performed significantly better under the Bilingual interface than under the Monolingual Spanish mode ($p < 0.05$).

However, there was considerable within-subject, between-task variability— a subject's performance on one task, as measured by F-measure, was only marginally well-correlated with their performance on another (see Figure 9.23 for scatterplots

⁸ For the sake of completeness, Table 9.26 gives the means of these values.

comparing individual subjects' performance under one mode to their performance under another). For example, the Pearson correlation coefficient (r) between subjects' F-measure scores under the Monolingual English interface ("Mono EN") and those under the Monolingual Spanish ("Mono ES"⁹) interface was 0.42 ($p \approx 0.001$; $r^2 = 0.17$). Correlation between subjects' performance under the Monolingual English and Bilingual interfaces was slightly better¹⁰, with a Pearson coefficient of 0.57 ($p < 0.001$; $r^2 = 0.32$) (indeed, Figure 9.23.C does have more of a visible trend than the other two plots). Performance under the Spanish-only and Bilingual correlated to a similar degree¹¹ as the English-only and Spanish-only correlations, with $r = 0.41$ ($p \approx 0.001$, $r^2 = 0.17185$).

Recall was somewhat better correlated within-subject and across-task, with a Mono EN/Mono ES r of 0.66 ($p < 0.001$), a Mono EN/Bi r of 0.69 ($p < 0.001$) and a Mono ES/Bi r of 0.56 ($p < 0.001$). However, Wolfe's test found that the three correlation coefficients were not statistically significantly different from one another,¹² suggesting that the amount of within-subject correlation was equivalent across all tasks. Since a subject's recall is heavily influenced by the number of documents they select, it is possible that the higher level of within-subject correlation that we see with recall (as compared with that seen with F-measure) may be related to the relatively high within-subject consistency of document selection count.

The subjects' precision scores tell a somewhat different story than that told by recall. Subjects precision scores were more consistent (Wolfe's test, two-sided $p < 0.001$) between the Mono EN and Mono ES interfaces ($r = 0.61$, $p < 0.001$) than

⁹We will be using "ES" as shorthand for "Español."

¹⁰ Wolfe's Test for Comparing Dependent Correlation Coefficients[210] tests the null hypothesis that $r(x, y) = r(x, z)$ against the null hypothesis that $r(x, y) \neq r(x, z)$, where r is the Pearson Correlation. In this case, the Wolfe test was used to compare the correlations between subjects' F-measures under the Monolingual English and Monolingual Spanish interfaces to the correlations between their F-measures from the Monolingual English and Bilingual interfaces, and found a p -value of 0.047, indicating that the two correlation coefficients were significantly different from one another.

¹¹ Wolfe's test, $p = 0.943$.

¹² EN/ES vs EN/BI: two-sided $p = 0.150$; ES/EN vs. ES/BI: two-sided $p = 0.526$.

between the Mono EN and Bilingual interfaces ($r = 0.45$, $p < 0.001$). However, their precision scores using the Mono ES interfaces were not at all correlated with their scores using the Bilingual interface ($r = 0.12$, $p = 0.349$).

English reading ability, as measured by the ILR instrument (see Section 8.1.2), did not seem to play a major role in determining subjects' performance on any of the tasks. Figure 9.24 on page 211 shows each ILR band's mean F-measure for each interface mode. The first thing to notice is that, consistently, subjects of all bands performed best under the Monolingual English interface mode. The second thing to notice is that any between-band performance difference is very minor at best, and that the highest band does not appear to have outperformed the other bands consistently. The final important thing to note in this figure is that there does not appear to be any major interaction between ILR level and interface mode: the overall pattern of performance is roughly the same across bands. The only exception to this trend is seen in the lowest ILR band, whose subjects seemed to perform better under the Monolingual Spanish task than they did under the Bilingual task, in contrast to the subjects in the other bands.

That said, consider Figure 9.25, which is the same as Figure 9.24 with the addition of error bars indicating 95% confidence intervals about each point. Note that, in all cases, each band's error bars overlap substantially, probably due to some bands' low sample sizes and high levels of variability within all bands. Given this finding, we are reluctant to make any conclusions regarding what effects a subject's language ability might or might not have their performance at the document selection task.

As was the case with selection counts, subject expertise and language proficiency were unassociated with F-measure for any of the interface modes.

Recall from Section 7.6.1 that the selection interface we provided to our subjects included both article titles as well as abstracts, but that the document surrogates

Figure 9.18: Histograms of the raw number of document selections made by subjects. Note the large amount of between-subject variation (i.e., some subjects selected many articles, some selected very few, and there were many in between).

displayed only the title by default. This design is quite standard for literature search engines, as including abstract text in with a lengthy list of results can lead to a difficult-to-use and cluttered interface.¹³ In order to actually view the abstract, subjects had to click on a link labeled (in either English or Spanish, as was appropriate for the interface) “abstract.”

We expected that this feature would see heavy use, since article abstracts often contain information that is necessary for making an accurate relevance judgment. To our surprise, however, only 14 of our 59 subjects looked at any abstracts at all, and only 11 of these looked at abstracts during the three experimental tasks (as opposed to the warmup or followup tasks). Of those 11, six looked at fewer than ten abstracts in total. The number of abstracts viewed by the remaining five subjects varied widely, from 13 to 90 abstracts.

Subjects who viewed abstracts did not have higher F-measures on any task than those who did not view abstracts (Kruskal-Wallis test, $p = 0.60$, $p = 0.73$, and $p = 0.79$ for the Monolingual English, Monolingual Spanish, and Bilingual interfaces, respectively).

¹³See Hearst’s “Search User Interfaces”[71] for further discussion of why and how one might display document surrogates.

Interface	Mean	Median	Std. Dev.
Mono EN	15.5085	16	7.76236
Bilingual	15.9322	16	9.04751
Mono ES	17.3559	19	8.68563

Table 9.24: Mean, median, and standard deviation for the number of articles selected under the various interface modes. Note the extremely high standard deviation; note further that subjects made more selections using the Monolingual Spanish interface than the other two.

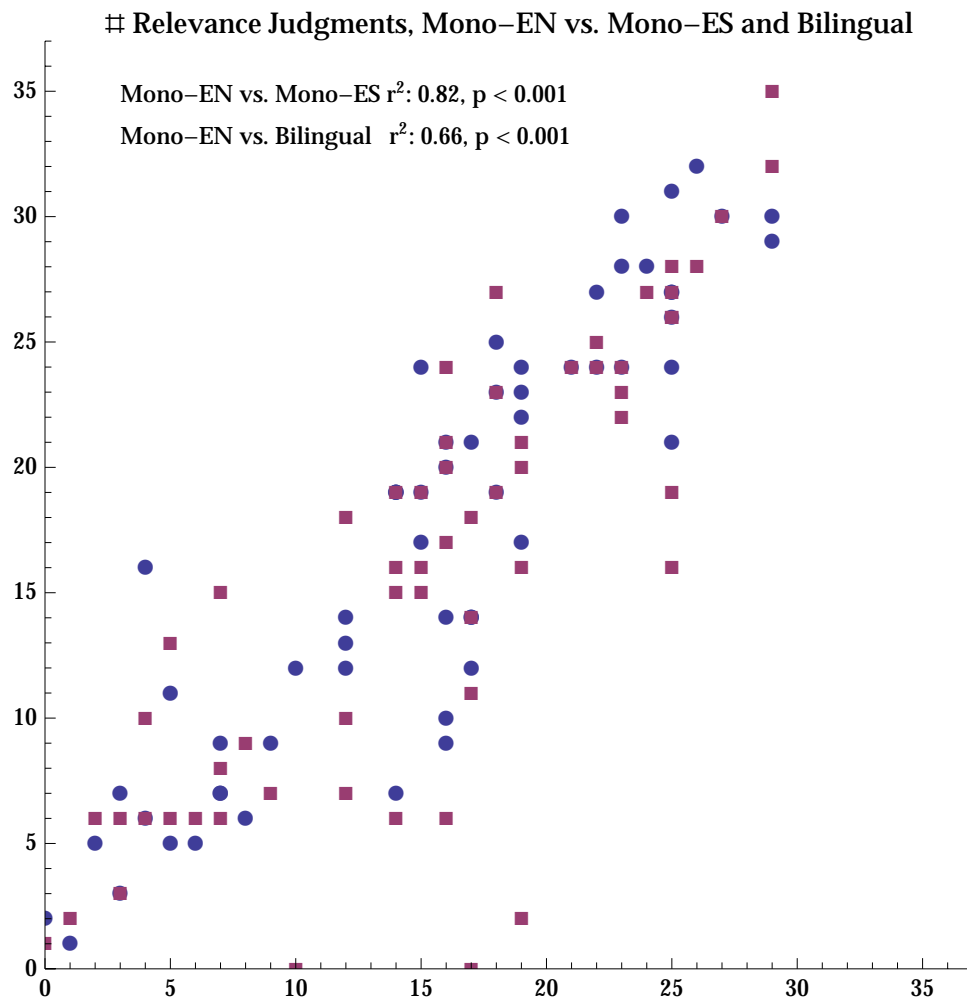


Figure 9.19: Subject selection counts from the English-only task plotted against counts from the Spanish-only and Bilingual tasks (circles and squares, respectively). Subjects were relatively self-consistent in terms of document selection counts; subjects who selected many articles under one mode were quite likely to select many articles under the other modes, and vice-versa.

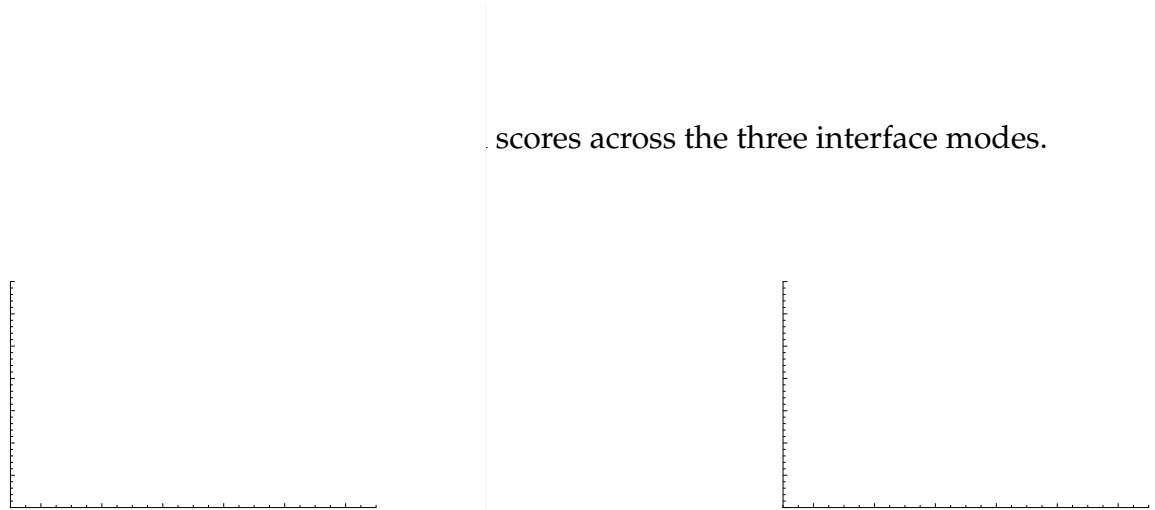


Figure 9.21: Distribution of recall scores across the three interface modes.

Figure 9.22: Distribution of F-measure scores across the three interface modes.

	En	Bi	Es
Precision	0.333 (0.198)	0.200 (0.121)	0.167 (0.165)
Recall	0.500 (0.217)	0.500 (0.284)	0.375 (0.206)
F-Measure	0.387 (0.151)	0.276 (0.148)	0.235 (0.118)

Table 9.25: Median precision, recall, and F-measure scores for each task. Standard deviations are in parentheses.

	En	Bi	Es
Precision	0.373 (0.198)	0.198 (0.121)	0.211 (0.165)
Recall	0.451 (0.217)	0.426 (0.284)	0.390 (0.206)
F-Measure	0.373 (0.151)	0.255 (0.148)	0.241 (0.118)

Table 9.26: Mean precision, recall, and F-measure scores for each task (Monolingual English, Bilingual, and Monolingual Spanish). Standard deviations for each value are in parentheses.

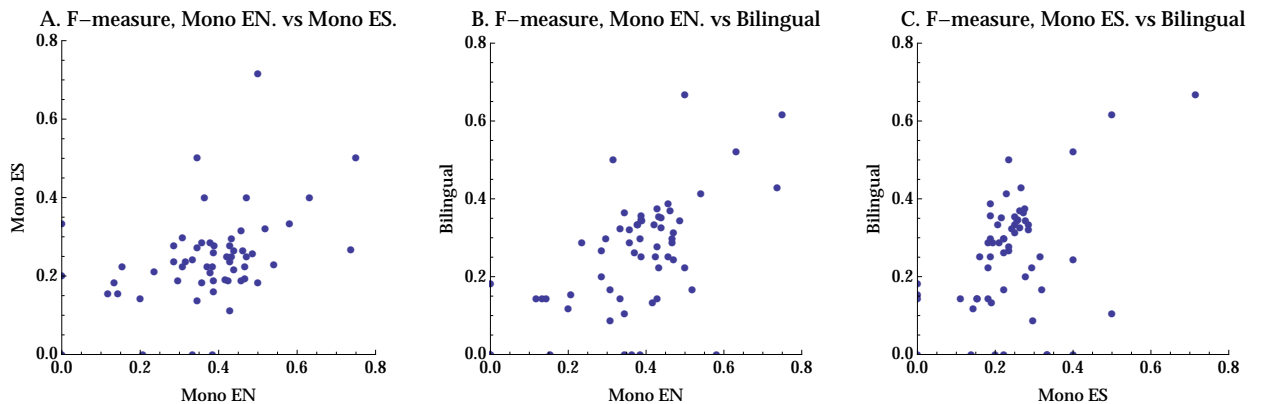


Figure 9.23: *A–B*: Scatterplots showing comparisons of the F-measures achieved by individual subjects when using the Monolingual English interface (abscissa) against their F-measures under the Monolingual Spanish interface (ordinate). *C* shows the equivalent comparison between the Monolingual Spanish interface F-measures and the Bilingual F-measures. Note that while there is a vague up-and-rightward cast to the data, the plots in general are quite noisy.

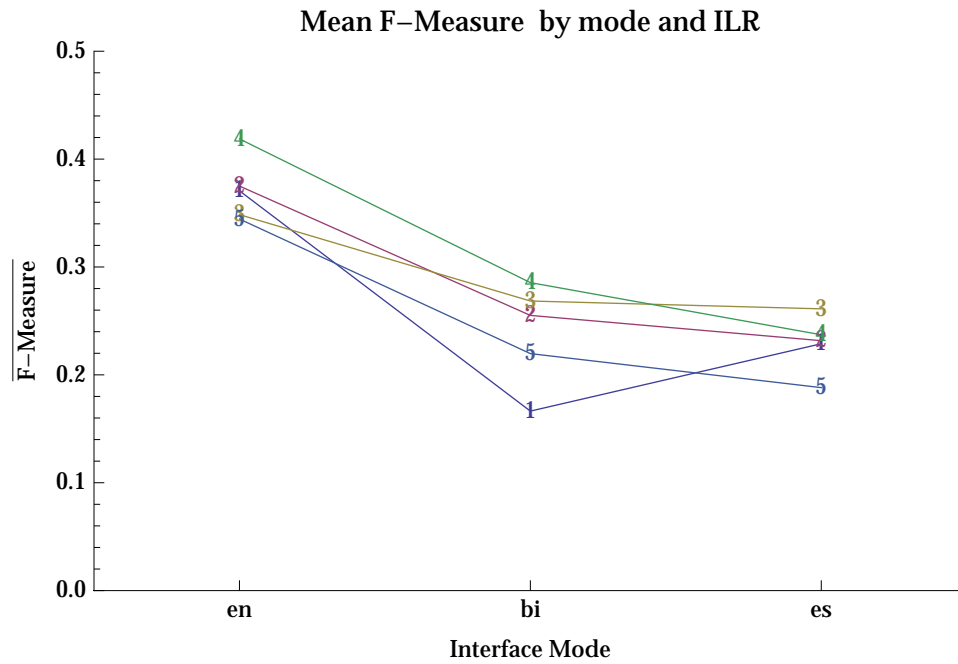


Figure 9.24: Mean F-measure for each task, stratified by ILR score.

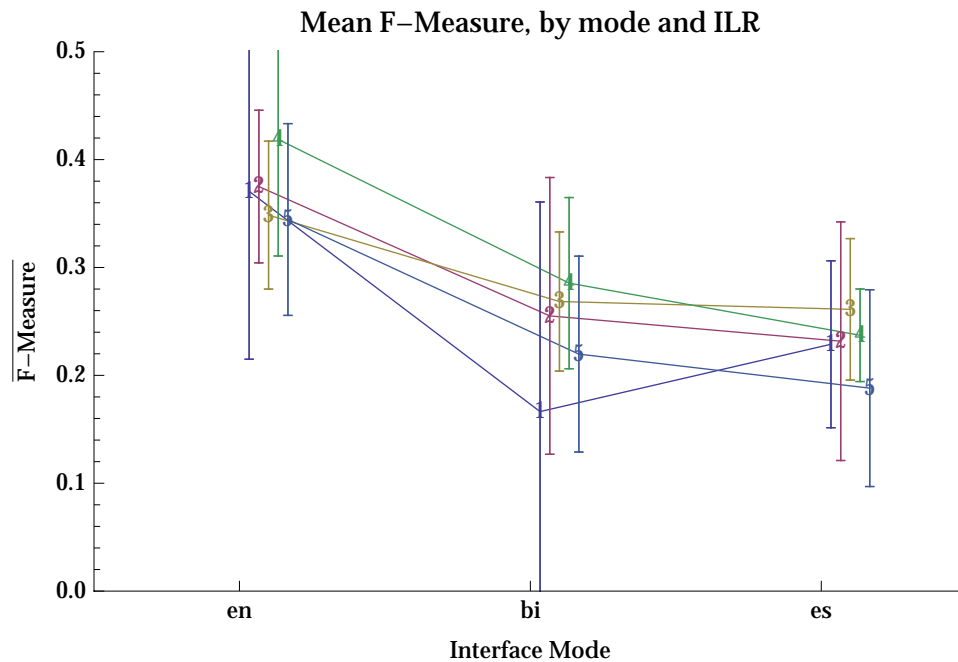


Figure 9.25: Mean F-measure for each task, stratified by ILR score, including error bars representing 95% confidence intervals about each data point. Note the wide confidence intervals, resulting from a combination of highly-variable data and low sample size.

9.4.2 Speed

As discussed in Section 8.3, our measure of subject speed for the document selection task was the elapsed time (in seconds) between the subject's first and last selection. As previously discussed, this is an imperfect method, but remains instructive. Table 9.27 gives the mean and median elapsed time for each task, along with standard deviation, and Figure 9.26 gives histograms for the elapsed times observed under each interface mode. There are two important points to note. First, the Monolingual English task has a much higher standard deviation than the other two tasks; this is due to the presence of a very small number of extreme outliers (i.e., subjects who took an abnormally long time to complete the task under this condition). Second, note that both the mean and median elapsed times are shortest for the Monolingual Spanish task; that said, the Kruskal-Wallis test finds no significant difference between the medians for the three tasks ($p = 0.278$).

Table 9.28 presents within-subject correlation in elapsed time between the various modes. At first glance, it appears as though the amount of time that subjects spent on one task was not particularly strongly correlated with how long they would spend on another. As previously mentioned, however, there were a small number of extreme outliers, particularly on the Monolingual English task. Table 9.29 gives the same data as Table 9.28, except with outliers (defined as subjects whose elapsed time was more than two standard deviations above the median) removed. Note that after removing a very small number of obvious outliers (three from each of both the Mono EN and Mono ES tasks, and two from the Bilingual task), the correlation coefficients increase dramatically, indicating a larger amount of within-subject consistency in terms of time spent on the selection task.

Another way to assess the relative amount of time that subjects spent using each interface would be to look at the *within-subject difference* in time spent between two modes; for example, looking at how much more (or less) time a sub-

	Mean	Median	Standard Deviation
Mono EN	366.04	258.93	400.10
Mono ES	258.04	220.35	163.83
Bilingual	290.85	273.01	174.77

Table 9.27: Mean, median, and standard deviation of elapsed time, by task, in seconds. The high standard deviation of times in the Monolingual English task is due to the presence of a very small number of extreme outliers.

ject spent completing the task when using the Monolingual Spanish interface than when using the Monolingual English interface. This can be done by subtracting each subject's elapsed time on one interface (interface "A") from their elapsed time on another interface (interface "B"); if the result is positive, they spent longer on interface A than they did on interface B.

Table 9.30 shows the median time differences between the various interfaces. It shows, for example, that subjects spent a median of 27.89 seconds *less* time using the Monolingual Spanish interface than they did using the Monolingual English interface, and that this difference was statistically significant (Wilcoxon Signed-Rank test, two-sided $p = 0.036$). Similarly, subjects spent a median of 45.69 seconds *less* time using the Monolingual Spanish interface than they did using the Bilingual interface to complete the document selection task ($p = 0.022$). This is in spite of the fact that subjects selected *more articles* using the Monolingual Spanish interface than the Monolingual English interface (see Section 9.4.1).

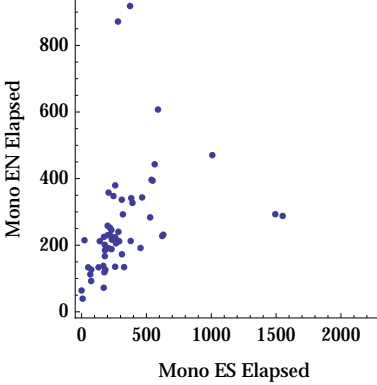
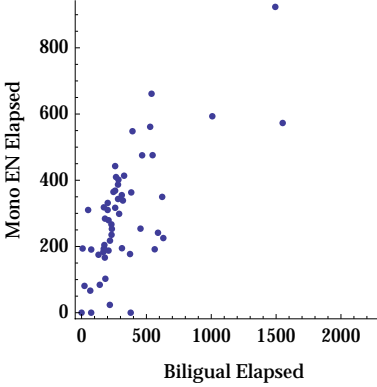
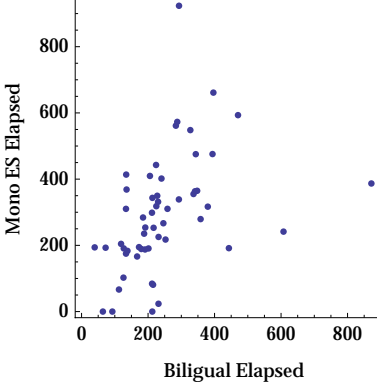
Mode	r		Scatterplot
Mono EN/Mono ES	0.20213	0.13520	
Mono EN/Biligual	0.45695	0.00039987	
Mono ES/Biligual	0.30347	0.022987	

Table 9.28: Within-subject correlation in elapsed time, between tasks.

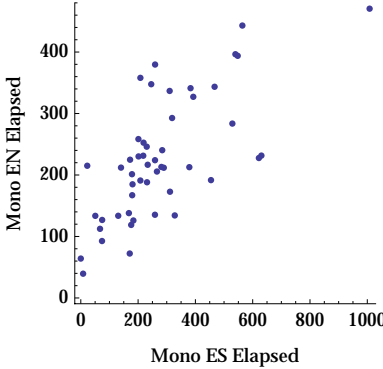
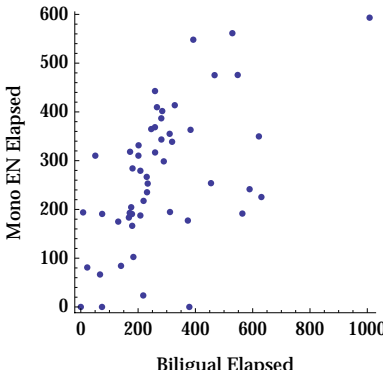
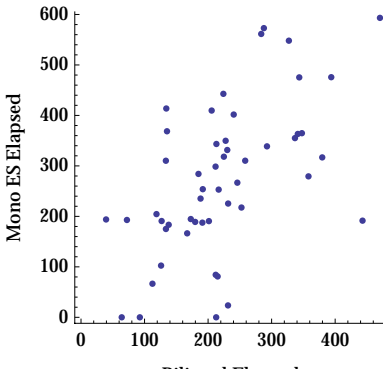
Mode	r	p	Scatterplot
Mono EN/Mono ES	0.69744	1.8248×10^{-8}	
Mono EN/Biligual	0.55496	0.000019584	
Mono ES/Biligual	0.57022	0.000012529	

Table 9.29: The same data as Table 9.28, but with outliers removed. Note the much higher Pearson correlation coefficients.

Figure 9.26: Histograms of elapsed time on the document selection task. Note the radically different abscissa on the Monolingual English histogram, resulting from a small number of extreme outliers.

Contrast	$\tilde{\Delta}t$ (seconds)	Interpretation	
Mono EN vs Mono Es	27.89	ES faster than EN	*
Mono EN vs Bilingual	-17.53	EN faster than Bilingual	
Mono ES vs Bilingual	-45.69	ES faster than Bilingual	*

Table 9.30: Median differences in elapsed time between interfaces, within subject. Rows marked with asterisks indicate that the difference between the medians of the time differences between two indicated interfaces was significant (Wilcoxon Signed-Rank test $p < 0.05$; H_0 : median time difference is 0).

Chapter 10

Discussion

The results of our evaluation were unsurprising in some ways and perplexing in others.

10.1 Preferences

Recall from Section 8.3 that our preference-related hypothesis had been that subjects would prefer the Bilingual interface mode to the other two. Figure 9.12 shows subject responses; for three of the four “positive” questions (in this case, “easiest,” “fastest,” and “liked best”), the Bilingual interface was indeed chosen by a majority of subjects. However, the difference majority was extremely small in absolute terms— between two and five percentage points. Fewer subjects found the Bilingual interface to be the “most difficult” one to use than either of the other two choices,¹ but again, the difference was not large. While we may certainly conclude that the Bilingual interface was a popular one among our subject population, our results do not allow us to definitively state that it was the overall “preferred” interface among our subject population.

¹The Monolingual English interface was by far the most “popular” one for subjects to mark as “most difficult.”

However, the fact that so many of our subjects did seem to appreciate the bilingual result presentation mode suggests that it is a feature worth exploring further, and that designers of these sorts of systems would do well to consider including bilingual results in their interfaces. When we expand the question, and compare how subjects felt about interfaces that contained *any* machine-translated Spanish content against how they felt about the Monolingual English interface, the picture begins to be clearer. Table 9.18 suggests that the majority of subjects preferred interfaces containing translated content than preferred interfaces to interfaces without such content; however, while framing the question in this manner does lead to differences that are somewhat larger, the sample sizes are still relatively small... to the extent that only one of the questions (the one asking subjects which interface they felt was fastest to use) had answers that were statistically different from a simple 50/50 split.

From this, then, we may conclude that while many (in fact, most) subjects found interfaces containing translated content to be superior in some way to English-only interfaces, but that this feeling was far from universal. One obvious question is what role English proficiency plays in subjects' opinions; logically, we would expect subjects with stronger English skills to benefit less from translated content than subjects with weaker skills, and to therefore be less likely to prefer interfaces that include translated content (or, at the very least, to be much more likely to prefer the Monolingual English interface).

Indeed, as discussed in Section 9.3, English reading proficiency played a major role in subjects' preferences. Figure 9.14 illustrates the "Some ES"/"Mono EN" dichotomized results shown in Table 9.18, stratified by ILR level. Note that the proportion of subjects preferring Spanish-containing interfaces drops steadily as ILR levels increase, demonstrating that the translated content was much more popular among subjects with lower English proficiency.

There is more to the story, however. Consider Figure 9.17, which illustrates the proportion of subjects who specified the Bilingual interface as the fastest, easiest, etc. vs. those who specified either of the two Monolingual interfaces. Note that while the proportion of subjects preferring the Bilingual mode does appear to be slightly higher at the lower ends of the ILR scale, there is nowhere near as obvious or strong of a trend as is seen when looking at the subjects who prefer one or the other of the monolingual modes (figures 9.15 and 9.16).

Figure 9.13 illustrates this finding nicely; note that subjects' responses to the four positively-worded questions (subfigures A–D) display a clear right-to-left diagonal pattern— subjects at the low end of the ILR scale very rarely said that the *English-only* mode was easiest, most efficient, etc., whereas subjects at the higher end of the scale rarely found the *Spanish-only* mode to be easiest, most efficient, etc. The Bilingual interface does not follow this pattern: a noticeable number of the lower-ILR subjects preferred the bilingual interface over the Monolingual Spanish interface, and a noticeable number of the high-ILR subjects preferred the Bilingual interface over the Monolingual English.

The exceptions to this pattern are instructive. Among the positively-worded UI feedback questions, the only one that did not follow this pattern (i.e., opposite behavior regarding the Monolingual Spanish interface between low-ILR and high-ILR subjects) was the “Fastest” question (Figure 9.13.C), to which a noticeable number of high-ILR subjects responded that they felt the Monolingual Spanish interface to be the fastest (note, however, that the inverse did not appear to be true: as was the case with all of the other positive questions, no low-ILR subjects found the Monolingual English interface to be fastest). The other exception to this pattern was with the negatively worded question (which asked which interface the subjects found most difficult to use), in response to which many high-ILR subjects reported finding the English-only interface to be the most difficult one to use.

We interpret these findings as follows. Clearly, the subjects who appreciated the translated content the most were those with the lowest English reading proficiency. However, this was far from universal, and it would be premature to “write off” the value of translated content to more proficient users. Even among our highest-ILR subjects, there was a sizable contingent that expressed preferences for the Bilingual interface mode, which suggests that they were deriving some benefit from translation support. Furthermore, many subjects in the higher ILR levels found the Monolingual English mode to be most difficult, and some number felt fastest at using the Monolingual Spanish mode. Among subjects who reported “liking” the Monolingual English mode the best, several reported that it was also the most “difficult.” Clearly, there exists room even among users who are very proficient in English for translation support.

One area that would benefit from further research is the relationship between subjects’ opinions of the Bilingual and Monolingual Spanish interfaces. When comparing subjects’ responses to one question to their responses to another, we observed some intriguing patterns (see tables 9.20 through 9.23). As previously discussed, there was a great deal of “cross-talk” between these two interface modes. For example, of the subjects who found the Monolingual Spanish interface to be “easiest,” as many “liked” the Bilingual interface best as “liked” the Monolingual Spanish interface best.

This pattern presented itself repeatedly among subjects who preferred one of the two Spanish-containing interfaces. We could interpret this finding in several ways: first, that these subjects simply did not differentiate between the two interfaces, and felt that they were of equal value; second, that the subjects *did* differentiate between the two, and did so for reasons that would be relevant to future designers of translated user interfaces. This strikes us as a situation that is ideally-suited for exploration using qualitative methods.

One open question is that of whether the quality of the machine-translated content affected user preferences. As described in Section 9.3, many of our subjects—even those that expressed a preference for translated content—complained about the variable quality of Google’s translations (as presented by our system). If the Spanish-language content was perfectly and fluently translated from English, would our higher-ILR subjects have still preferred the Monolingual English interface? We discuss additional future research questions in chapter 11.

10.2 Accuracy

As described in Section 9.4.1, our subjects seemed to perform best at the classification task while using the Monolingual English interface. We had initially hypothesized that subjects would perform best under the bilingual interface mode; this does not appear to have been the case. However, as shown in Table 9.25 and figures 9.20–9.22, this observation does not necessarily say very much about the relative performance of subjects under different interfaces, as our subjects did not perform particularly well at the document selection task under *any* of the interfaces. Under the Monolingual English interface mode, most subjects achieved F-measures somewhere between 0.3 and 0.5. Under the two modes using translated content, matters were even worse: under the Monolingual Spanish mode, the vast majority came in at less than 0.3, and using the Bilingual mode, virtually none were able to break 0.4. Furthermore, as shown in Figure 9.21, subjects’ recall scores appeared to be nearly random, and as shown in Figure 9.20, the precisions were even lower than the F-measures.

Figure 10.1 shows histograms of the false positive rates² exhibited by our subjects. Note that the false positive rates are quite high. What could have caused this?

² Defined as $fp/(tp + fp)$, where “fp” is the number of false positive classifications made by a subject and “tp” is the number of true-positive classifications.

Figure 10.1: False positive rates, by task.

Recall from Section 8.2.1 that approximately 25% of the articles in each document subset were relevant, and that subjects interacted with those articles in groups of forty articles per interface. This means that, under any given interface, approximately ten of the forty articles would have been relevant, and the rest not relevant. From Figure 9.18 we see that the majority of subjects selected more than ten articles. This explains the low precision and high false positive rates, but raises the question of *why* subjects selected so many more articles than were actually “relevant.”

Examining within-subject performance data raised additional questions. As previously discussed, a subject’s recall when using one interface was moderately well-correlated with their recall when using the other two interfaces. This was unsurprising, given that selection count was quite well-correlated within-subject, as shown in Figure 9.19. However, precision and F-measure were not nearly so well-correlated, indicating that subjects’ accuracy varied widely from task to task. Some variation in accuracy was to be expected, due both to the fact that the article sets (while all drawn from the same pool of articles, and randomly partitioned) were different from interface to interface, and, of course, the possibility that the interface itself would affect the subjects’ accuracy. The level of within-subject variation that we saw, however, struck us as more than could be explained by either cause.

When we designed the task, we anticipated that it would be a relatively simple

Figure 10.2: Selection count, by article. Each bar represents an article; its height, the number of times it was marked as relevant by our subjects; they are sorted by selection frequency, *not* the order in which they were presented to subjects.

one for subjects to carry out, and our initial pilot testing indicated to us that, indeed, our subjects were having a relatively easy time identifying relevant articles. Our subjects' low performance and high within-subject variation led us to think that we had perhaps been mistaken on this count.

To attempt to understand what caused this low performance and high variability, we began to look into the individual subjects' selections (i.e., which specific articles were selected as "relevant" by which individual subjects) from each interface mode. Figure 10.2 shows the selection frequency for each article, by interface mode. Each bar represents a single article; its height, the number of times it was marked as relevant by our subjects. The bars/articles are sorted by selection frequency, *not* the order in which they were presented to subjects. The large variation in terms of selection count suggests that subjects were not selecting at random; some articles were marked as relevant by almost all subjects, others hardly ever. Had subjects indeed been making their selections completely at random, we would expect a much flatter profile, with smaller differences between more and less "popular" articles.

From this, we concluded that our subjects' low performance was not due to a lack of trying on their part— they were clearly doing *something*, as opposed to simply clicking randomly. Perhaps the task was indeed more difficult than we

thought, though, and subjects were having trouble identifying the relevant articles. We began to look into differences between articles that were selected frequently and articles that were not selected frequently. Our first thought was to examine the distribution of “relevant” articles in terms of selection frequency. We expected to find that relevant articles would be selected more often than non-relevant articles— even with the high number of false positives identified by our subjects, we assumed that they would have tended to recognize the relevant articles even as they selected additional ones.

Figure 10.3 shows the same selection frequency distribution as Figure 10.2, with each bar color-coded to indicate its relevance status according to the original ground truth used to compute the various performance metrics. Green bars are relevant; black bars are not. We expected to see the green bars clustered together toward the left-hand side of each graph— in other words, we expected the relevant articles to consistently be among those that were frequently selected by subjects. As shown in this figure, this was not the case. The green bars (relevant articles) are scattered throughout the relevance range; some of the relevant articles were indeed selected with greater frequency, but just as many were selected much more rarely. In short, an article’s ground-truth relevance appeared to have no relationship to the frequency with which it was marked as relevant by our subjects.

We arrived at two possible explanations for this finding:

1. Our subjects were performing the task as instructed, but our ground truth was flawed, thereby causing us to incorrectly measure their performance;
2. Our subjects were *not* performing the task as instructed, and were instead using some other set (or sets) of criteria to make what they saw as valid relevance judgments.

Figure 10.3: The same figure as shown in Figure 10.2, with the addition of color-coding to indicate ground-truth relevance. Green bars represent articles that were relevant according to the Haynes queries; black bars were not relevant. Note that the distribution of green bars (relevant articles) is essentially even throughout the selection frequency range.

10.2.1 Ground Truth Assessment

In order to investigate the first explanation, we began looking at the specific articles that were frequently marked as relevant, as well as those that were less frequently selected by subjects. While we had previously examined the contents of our article collection, we had not done so with this level of detail. What we saw was troubling. The first set of problems we observed was with the articles themselves. Recall from Section 8.2.1 that we used the PubMed Clinical Queries to semi-automatically build our document collection. When we were initially validating the collection, the quality of the queries' results had seemed to us to be quite good; upon closer inspection, however, we discovered that the queries we used were insufficiently precise, and had retrieved a far broader variety of articles than we had previously thought.

Due to the way in which brain trauma articles are indexed in MEDLINE, our query retrieved several articles about diagnosing or treating brain injuries resulting from radiation treatment (e.g., "Multivoxel 3D proton MR spectroscopy in the distinction of recurrent glioma from radiation injury," by Zeng, et al.[211]), as well as several articles that were only nominally related to traumatic brain injury ("Meta-analysis of apparent diffusion coefficients in the newborn brain," by Coats,

et al.[212]). Additionally, we found a number of articles that, while clearly related to TBI, and perhaps even related to *diagnosing* TBI, were “abstract” enough that a non-expert clinician would be easily deceived. Most of these involved research articles from molecular biologists investigating biomarkers of TBI.

One example of such an article is “Cerebral apoptosis in severe traumatic brain injury patients: an in vitro, in vivo, and postmortem study,” by Miñambres et al.[213], which “aimed to analyze the presence of apoptosis and the expression of apoptosis-related proteins in brain samples from patients with TBI.” While the ultimate objective of this work was indeed to attempt to identify novel biomarkers that could potentially assist in diagnosing TBI, this connection is somewhat abstruse, and could easily be missed by a clinician without expertise in molecular biology. When we were initially assembling the collection, we knew that the Clinical Queries had pulled in several of these sorts of articles; however, we did not fully appreciate just *how many* had been included. In any given subset of 40 articles, there appear to be anywhere from three to seven articles whose appropriateness for inclusion would be debatable.

To make matters worse, some of these spurious articles were, according to our ground truth, “relevant.” Subjects who, correctly, did not mark the “Multivoxel...” article as being relevant to diagnosing traumatic brain injury were penalized in terms of their accuracy scores. This finding, together with the larger-than-expected heterogeneity within our document collection, led us to question the validity of the classification accuracy scores we had calculated.

When we initially validated the document collection, our raters were medical researchers with significant domain expertise, familiarity with the TBI literature, and experience at systematically reviewing collections of articles. As such, when asked to manually classify the articles, they agreed substantially with the classifications made by the Clinical Queries, and were able to intuit an understanding

of how to fit the less-obvious articles into the article set. In retrospect, it seems as though our validators were quite different from our subjects, and as such were perhaps not the best judges of what was relevant and what was not. If we had used validators who were *not* subject experts— who were, perhaps, closer to “average clinicians”— perhaps we would have discovered these problems sooner, and rejected the primarily unsupervised use of the Clinical Queries as a methodology for constructing our document collection.

To investigate this possibility, and to attempt to salvage the document selection data collected from our subjects, we enlisted the help of three Portland-based physicians to review all 200 of the articles in our document collection, and instructed them to identify any articles that were about traumatic brain injury *diagnosis*. Our thinking was that if these secondary raters produced similar results to the original judgments, we could safely conclude that our ground truth was relatively valid; if they produced different results, but were consistent within themselves, we could perhaps use their judgments as a new ground truth and recompute the accuracy statistics.

We gave our secondary raters the same instructions that we gave our subjects, and that we gave our previous group of human raters; however, due to time constraints, they did not use our document selection interface. Instead, we provided them with a Microsoft Excel³ spreadsheet containing the titles, and asked them to use that file to record their judgments. The secondary raters did not know how many “relevant” articles the set contained (i.e., we did not inform them of the nominal 25%/75% split), and they were also unaware of how the articles were split up among tasks.⁴

³Microsoft Corp., Redmond, WA.

⁴ For reasons that will be explained shortly, we also asked them to identify articles that they felt were simply “about” traumatic brain injury in general (in addition to identifying articles specifically about diagnosing TBI). Until noted otherwise, all figures and tables refer to the secondary raters’ judgments regarding which articles were specifically about diagnosing TBI.

The results of this secondary rating were instructive. First and foremost, our second group of raters reported a much greater amount of uncertainty and ambiguity than did our initial raters, indicating that the task was not as simple or clear-cut for non-expert clinicians as we had originally thought. The second instructive finding from the secondary raters was how different each rater's judgments were from the others. One rater (rater "B" in the tables and figures below) was extremely conservative, and identified very few articles as being about TBI diagnosis, while another rater (rater "C") was far more permissive and identified nearly half of all of the articles as being about TBI diagnosis. The third rater ("A") was somewhere in between, but leaned towards the conservative side. Figure 10.4 shows graphs similar to those in Figure 10.3, but with the highlighted bars representing articles marked as relevant by raters A, B, and C, as appropriate. Note that rater C's selections were much closer to those made by many of our subjects than those made by raters A and B. In any event, while there was some overlap between the judgments made by our secondary raters and the original ground truth, there were substantial differences, as shown in Table 10.1.

Table 10.2 shows Cohen's Kappa scores between the three raters and the original judgments (marked "orig" in the table). Note the relatively low levels of agreement between the raters and the original judgments⁵, and between raters A and B and rater C. Raters A and B had relatively high levels of agreement with one another, largely because rater B's judgments were essentially a subset of those made by rater A (as can be seen in Figure 10.4).

Recall that one of our goals in obtaining new relevance judgments was to be able to make use of our subjects' document selections to investigate the effects of interface mode on selection performance. Using the original relevance judgments, we had found that (contrary to our expectations) subjects seemed to perform better,

⁵Using the rubric from [214], as described by [204].

on average, when using the monolingual English interface mode, and the differences in performance between modes were relatively large. Of course, due to the problems with the original ground truth, this finding is suspect. If, however, we observed similar results when using the secondary raters' judgments as ground truth, we could perhaps conclude that the accuracy findings were not completely invalid.

When we re-computed our various accuracy measures using the secondary raters' judgments, we saw a different picture than we had seen when using the original judgments. Given the low levels of agreement between the secondary raters, we re-computed F-measure three times—once using each secondary rater's set of judgements. The results are shown in Table 10.3 on page 232; note first that each rater's judgements led to widely varying F-measure calculations. Also note that the overall pattern we saw with the original judgments—subjects performing best when using the monolingual English interface, followed by the bilingual and then the monolingual Spanish interfaces, but with relatively minor differences—did not persist when we used the secondary raters' judgements.

When using rater A's judgments, subjects appeared to perform quite a bit better under the bilingual and monolingual Spanish interfaces than under the monolingual English interface; when using rater B's judgments, we see a similar trend, although the differences are not as great and the general numbers are much lower. Using rater C's judgments, we see an entirely different pattern. First, subjects appeared to perform quite a bit better than under the other two secondary raters' judgments; second, subjects appeared to perform best under the monolingual English interface, followed closely by the monolingual Spanish task and then the bilingual task. The first observation—that subjects seemed to perform better when judged using rater C's judgments as a ground truth—is unsurprising, given how much closer rater C's judgments were to those made by our subjects (see Fig-

accuracy scores might mean, it must be pointed out that the point is largely moot, since the secondary relevance judgements are unusable in their current form. The viewer of Figure 10.4 will have noticed that the distribution of “relevant” articles from interface to interface (as determined by the alternate raters) is extremely uneven; each interface has a very different number of “relevant” articles. Our task was designed for a 25%/75% relevant/not-relevant distribution of articles in each interface’s subset; barring that, the task required a roughly equivalent number of relevant articles in each interface’s subset. If one interface had more relevant articles than another, that could strongly affect subject behavior.

The relevance judgments made by rater A were very unevenly distributed between the various interface modes (see Table 10.4); rater B’s selections happened to be reasonably evenly distributed across modes, but B marked as relevant far too few articles to be useful. Rater C, however, selected a reasonable number of articles, and those selections happened to be distributed relatively evenly across interface modes. However, despite this (and other) encouraging attributes to rater C’s judgments, we felt that there was sufficient variation among our raters to preclude our using any single rater’s judgments as a gold standard, and we are therefore hesitant to make use of the subject performance data calculated using those judgments.

Furthermore, recall that earlier we mentioned not one but *two* possible explanations for the our subjects’ perplexingly low accuracy scores. The first explanation was that our subjects had been performing the task as instructed, but our calculations were incorrect due to flawed ground truth. The second explanation was that our subjects (or a sizable proportion thereof) were *not* performing the task as instructed, and were instead using some other set of criteria to determine which articles to select. Our discussion now turns to this second explanation.

	O +	O -		O +	O -		O +	O -
A +	15	13	B +	9	4	C +	13	27
A -	13	79	B -	19	88	C -	15	65

Table 10.1: 2×2 tables comparing the judgments made by raters A, B, and C against the original ground truth (O). See Table 10.2 for Kappa scores.

orig	0.449664	0.480315	0.226994
0.449664	'A' Diag	0.649514	0.408357
0.480315	0.649514	'B' Diag	0.397178
0.226994	0.408357	0.397178	'C' Diag

Table 10.2: Cohen's Kappa values between the original, Clinical Queries-derived ground truth ("orig") and the three secondary raters. Note that raters A and B have a "fair" kappa with one another, according to the rubric described by [204], while their kappas with the original judgments are lower. Rater C has a particularly low level of agreement with the original judgments, and questionable levels of agreement with the other two raters.

Rater	Mono EN	Bilingual	Mono ES
A	0.2759 (0.1485)	0.5000 (0.1933)	0.4348 (0.1247)
B	0.2353 (0.1440)	0.2857 (0.1243)	0.3077 (0.1205)
C	0.6190 (0.1818)	0.5128 (0.1704)	0.5806 (0.1943)

Table 10.3: Median F-measures (with standard deviations) for each interface mode, using judgments from each of the secondary raters. Compare to Table 9.25.

Rater	Mono EN	Mono ES	Bilingual
A	4	9	15
B	3	5	5
C	13	15	12

Table 10.4: Counts of "relevant" articles by secondary raters, by interface mode.

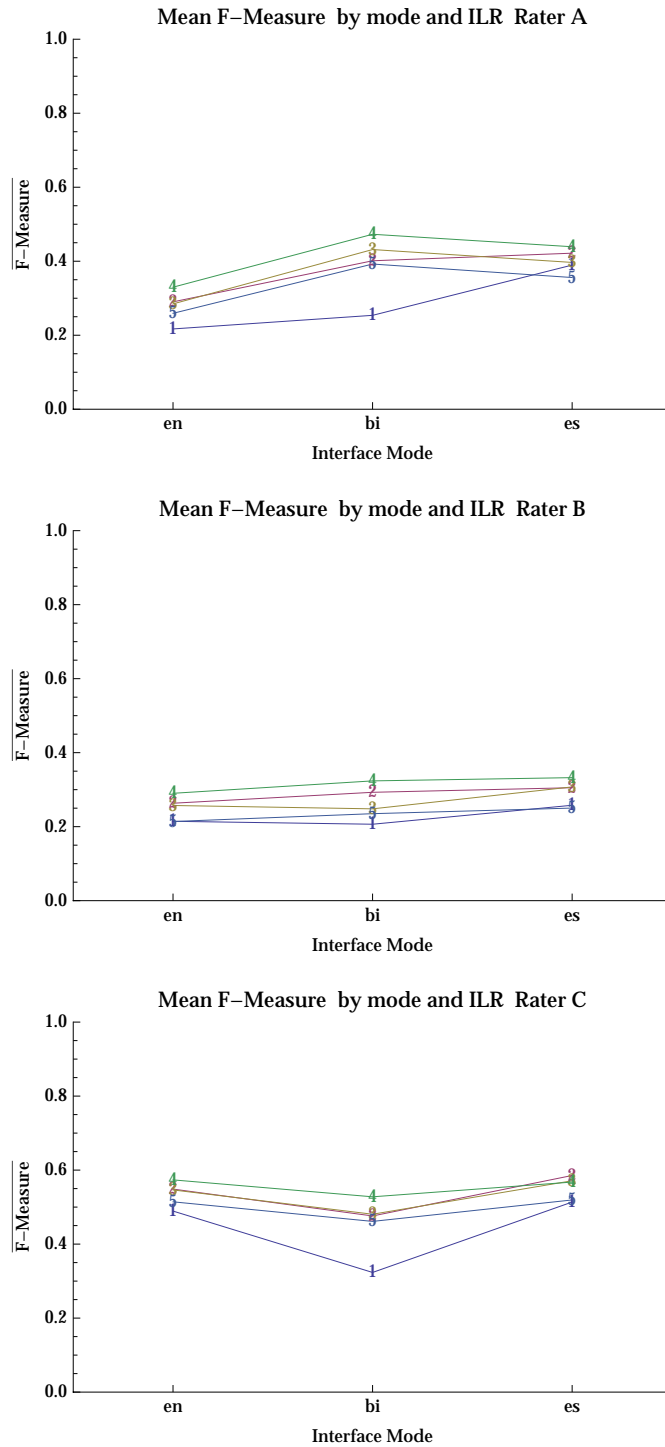


Figure 10.5: Mean F-measure, by interface mode and ILR, using the secondary raters' judgments. Compare with Figure 9.24.

10.2.2 A Different Task?

While the first explanation—flawed ground truth—did, as we saw, have some merit, we noticed some trends while reviewing our subjects' individual selections that led us to believe that perhaps it was not the end of the story. Consider Table 10.5 on page 238, which lists the titles⁶ of the ten most frequently-selected articles from the monolingual English interface mode's subset. Recall that subjects were instructed to select articles that they felt were about *diagnosing* traumatic brain injury, its side effects, or sequelae. Many of the most-commonly-selected articles, however, are emphatically *not* about diagnosing TBI: "Randomized *treatment* trial in mild traumatic brain injury," "Does intensive rehabilitation improve the functional outcome of patients with traumatic brain injury (TBI)? A *randomized controlled trial*," "Multicenter *trial* of early hypothermia in severe brain injury," etc. (emphasis ours). This same pattern held under the other two interface modes (tables 10.6 and 10.7).

We next examined the *least-commonly-selected* articles, in an attempt to see if there were any similar trends. Tables 10.8, 10.9, and 10.10 (pages 241–243) list the ten least-frequently-selected articles for the Monolingual English, Monolingual Spanish, and Bilingual interface modes, respectively. The first thing we noticed was that there were fewer obvious misclassifications—there seemed to be relatively few of what we would consider to be false negatives (i.e., articles about diagnosing TBI that our subjects considered to be not-relevant). In fact, many of the less-commonly-selected articles fell into the previously-mentioned category of articles that probably did not belong in our document collection in the first place, as many of them had connections to traumatic brain injury in general that were ten-

⁶ Since, as previously discussed, the vast majority of our subjects relied entirely on titles for their document selections, we decided to follow suit; after all, if our goal was to attempt to ascertain whether our subjects were using a different set of relevance criteria, we would need to attempt to view the documents in the same manner in which our subjects viewed them.

Figure 10.6: Article selection count, by interface mode, color-coded according to the “Title contains TBI” rubric. Green bars represent articles whose titles include the phrase “traumatic brain injury” or “traumatic brain injuries.” Note the extremely high degree of overlap between those articles most often selected by subjects and those whose titles mention TBI.

uous at best. After reviewing in detail the titles and their corresponding selection frequencies, we noticed that the articles that our subjects tended to select most frequently all seemed to have one thing in common: their titles explicitly mentioned (i.e., contained the words) traumatic brain injury.

To demonstrate this further, consider Figure 10.6, which is a version of Figure 10.3 using a novel set of relevance judgments generated according to the following rubric: articles whose titles contained the phrase “traumatic brain injury” or “traumatic brain injuries” were considered relevant, and articles without either of those phrases in their titles were considered to be not relevant. Note the almost perfect overlap between the green (relevant) articles and the most-frequently-selected articles. Under all three interface modes, the selection counts for articles whose titles contained these key phrases were significantly higher than the counts for articles whose titles did not mention traumatic brain injury (Student’s T-test, one-sided $p < 0.0001$, repeated for all three interface modes).

Unfortunately, we were unable to follow up with any of our subjects and ask them what it was that they thought they were doing when they were carrying out the document selection task. We believe, however, that the evidence suggests that our subjects (or at least, a substantial number of our subjects) were selecting the

articles that they felt to be about traumatic brain injury in *general* (as opposed to articles specifically about *diagnosing* traumatic brain injury, as the task's instructions had stated), and that they were basing this judgment largely on the word content of the titles. This interpretation is bolstered by the following statement from one of our subjects, in response to a question asking what they found easiest about the Monolingual English mode:

“terminos comunes como brain injury o head injury faciles de reconocer en los titulos, hacian mas rapida la seleccion del articulo” (Translation: Common terms, such as “brain injury” or “head injury” are easier to recognize in the titles, which makes selecting the article much faster)

When we re-computed classification accuracy statistics using this set of “relevance judgments,” we found that our subjects' performance appeared to increase substantially over what we saw when using the original set of relevance judgments, particularly in terms of F-measure (see Table 10.11, page 244). Interestingly, however, we did not see the same pattern of differences in performance between interface modes that we previously saw. While the Friedman test did find a difference between subjects' F-measures under the three interfaces ($p = 0.022$), post-hoc analysis using the Wilcoxon Sign-Rank test found that only the Monolingual Spanish and Bilingual interfaces differed significantly ($p = 0.009$), and that difference was very small (median difference in F-measure: 0.043).

We also did not observe any difference between ILR levels in terms of performance (see Figure 10.7)—with the exception of the subjects in the lowest ILR level, who appeared at first glance to have done considerably worse when using the Bilingual interface than when using the other two interfaces. Upon further inspection, however, we concluded that this was most likely an artifact of sample size (see Figure 10.8, which adds 95% confidence intervals to the graph shown in Figure 10.7—note the extremely wide intervals around all points, but particularly

around the point representing the mean score of subjects in the lowest ILR band using the Bilingual interface mode). It is unsurprising that we saw no real differences between ILR levels; if, in fact, the subjects were simply identifying articles whose titles explicitly mentioned TBI, the task would in essence be a relatively simple word-recognition problem, and English reading proficiency would play a smaller role in determining a subject's performance than if the task were one that required reading comprehension.

We cannot be at all sure about what our subjects thought they were doing while performing the document selection task. True, many of them *appear* to have been attempting to identify articles whose titles mentioned traumatic brain injury, but we cannot be sure, and there is also no way to tell whether some subset of our subjects were doing something else entirely. This precludes us from drawing any meaningful conclusions from these data, just as the serious and numerous issues with our document collection, discussed earlier in this chapter, prevented us from making use of the original ground truth. We are, in fact, unable to draw any conclusions whatsoever about the effects (if any) of interface mode on our subjects' ability to identify relevant articles. In Section 10.4, we discuss several possible explanations as to how this state of affairs came to be, and what we might do in the future to prevent it from happening again.

PMID	Title	Relevant	Count
16786351	Impaired cognitive functions in mild traumatic brain injury patients with normal and pathologic magnetic resonance imaging.	Y	49
16880017	Randomized treatment trial in mild traumatic brain injury.	N	46
19608224	Transcranial Doppler pulsatility index is not a reliable indicator of intracranial pressure in children with severe traumatic brain injury.	Y	46
17653942	Does intensive rehabilitation improve the functional outcome of patients with traumatic brain injury (TBI)? A randomized controlled trial.	N	44
18807008	Sodium lactate versus mannitol in the treatment of intracranial hypertensive episodes in severe traumatic brain-injured patients.	N	44
17651596	Effect of mild hypothermia on glucose metabolism and glycerol of brain tissue in patients with severe traumatic brain injury.	N	44
19858967	Emergency department assessment of mild traumatic brain injury and the prediction of postconcussive symptoms: a 3-month prospective study.	Y	43
19245306	Multicenter trial of early hypothermia in severe brain injury.	N	42
18363508	Cerebral apoptosis in severe traumatic brain injury patients: an in vitro, in vivo, and postmortem study.	Y	41
16299192	Derivation of a clinical decision rule to guide the inter-hospital transfer of patients with blunt traumatic brain injury.	Y	38

Table 10.5: The ten most commonly-selected articles under the Monolingual English interface mode. The first columns are as follows: the article’s PubMed identifier; the article’s title; whether or not the article was “relevant” under the original ground truth; the number of subjects who marked the article as “relevant.”

PMID	Title	Relevant	Count
19249933	Discrete cerebral hypothermia in the management of traumatic brain injury: a randomized controlled trial.	N	44
17638640	Prognostic study of using different monitoring modalities in treating severe traumatic brain injury.	N	44
18759980	Pentobarbital versus thiopental in the treatment of refractory intracranial hypertension in patients with traumatic brain injury: a randomized controlled trial.	N	43
17053267	Effect of continuous display of cerebral perfusion pressure on outcomes in patients with traumatic brain injury.	N	43
19388278	Cerebrovascular reactivity and autonomic drive following traumatic brain injury.	Y	43
17252176	Mild traumatic brain injuries: the impact of early intervention on late sequelae. A randomized controlled trial.	N	43
19779323	Erythropoiesis stimulating agent administration improves survival after severe traumatic brain injury: a matched case control study.	N	43
19257803	Validation of serum markers for blood-brain barrier disruption in traumatic brain injury.	Y	42
17853130	Using the Wechsler Memory Scale-III to detect malingering in mild traumatic brain injury.	Y	38
16966534	Effects of rivastigmine on cognitive function in patients with traumatic brain injury.	N	38

Table 10.6: The ten most commonly-selected articles under the Monolingual Spanish interface mode.

PMID	Title	Relevant	Count
18684008	Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics.	N	45
18355159	Hypothermia treatment for traumatic brain injury: a systematic review and meta-analysis.	N	42
19061735	A randomized controlled trial of holistic neuropsychologic rehabilitation after traumatic brain injury.	N	41
16304487	Validity of the Patient Health Questionnaire-9 in assessing depression following traumatic brain injury.	Y	40
19899831	Construct validity of an attention rating scale for traumatic brain injury.	N	39
19602473	S-100B and neuron specific enolase are poor outcome predictors in severe traumatic brain injury treated by an intracranial pressure targeted therapy.	N	38
19191091	Long-term effects of rivastigmine capsules in patients with traumatic brain injury.	N	38
18824999	Importance of screening logs in clinical trials for severe traumatic brain injury.	N	38
19345782	Sensitivity and specificity of the Beck Depression Inventory-II in persons with traumatic brain injury.	Y	37
18760149	Branched-chain amino acids may improve recovery from a vegetative or minimally conscious state in patients with traumatic brain injury: a pilot study.	N	37

Table 10.7: The ten most commonly-selected articles under the Bilingual interface mode.

PMID	Title	Relevant	Count
18053002	Prevention of traumatic headache, dizziness and fatigue with creatine administration. A pilot study.	N	6
17703885	Pregabalin in patients with central neuropathic pain: a randomized, double-blind, placebo-controlled trial of a flexible-dose regimen.	N	5
19255028	Effectiveness of educational materials designed to change knowledge and behaviors regarding crying and shaken-baby syndrome in mothers of newborns: a randomized, controlled trial.	N	4
18729534	Approved and investigational uses of modafinil : an evidence-based review.	N	4
18422412	Lessons learned: the effect of prior technology use on Web-based interventions.	N	4
17510596	Using interactive multimedia to teach parent advocacy skills: an exploratory study.	N	2
19255065	Do educational materials change knowledge and behaviour about crying and shaken baby syndrome? A randomized controlled trial.	N	2
15271412	Detecting symptom- and test-coached simulators with the test of memory malingering.	Y	2
19748046	Meta-analysis of apparent diffusion coefficients in the newborn brain.	Y	2
19620954	Incidence of clinically significant responses to zolpidem among patients with disorders of consciousness: a preliminary placebo controlled trial.	N	1

Table 10.8: The ten *least* commonly-selected articles under the Monolingual English interface mode.

PMID	Title	Relevant	Count
17414000	Effectiveness of an intravascular cooling method compared with a conventional cooling technique in neurologic patients.	N	15
17893578	Effects of fentanyl and S(+)-ketamine on cerebral hemodynamics, gastrointestinal motility, and need of vasopressors in patients with intracranial pathologies: a pilot study.	N	13
18467882	Developing alternative strategies for the treatment of traumatic haemorrhagic shock.	N	12
20220416	Brain death confirmation: comparison of computed tomographic angiography with nuclear medicine perfusion scan.	Y	12
19006900	Testing educational strategies for Shaken Baby Syndrome.	N	7
19801201	Neurocognition in patients with brain metastases treated with radiosurgery or radiosurgery plus whole-brain irradiation: a randomised controlled trial.	N	3
18556361	Medulla oblongata volume: a biomarker of spinal cord damage and disability in multiple sclerosis.	N	2
18609339	Malingering Scales for the Continuous Recognition Memory Test and the Continuous Visual Memory Test.	Y	2
17766155	Diagnosis of adult GH deficiency.	Y	1
19023743	A survey of neuropsychologists' practices and perspectives regarding the assessment of judgment ability.	N	0

Table 10.9: The ten *least* commonly-selected articles under the Monolingual Spanish interface mode.

PMID	Title	Relevant	Count
18660565	Clinical assessment of the HELLODOC tele-rehabilitation service.	N	10
17724554	Application and validation of the barrow neurological institute screen for higher cerebral functions in a control population and in patient groups commonly seen in neurorehabilitation.	Y	7
18056855	MR imaging of the brain 1 year after aneurysmal sub-arachnoid hemorrhage: randomized study comparing surgical with endovascular treatment.	N	7
18331207	Fetal electrocardiographic monitoring during labor in relation to cord blood levels of the brain-injury marker protein S-100.	N	7
18045737	Core temperature cooling in healthy volunteers after rapid intravenous infusion of cold and room temperature saline solution.	N	4
16757720	Stereotactic radiosurgery plus whole-brain radiation therapy vs stereotactic radiosurgery alone for treatment of brain metastases: a randomized controlled trial.	N	4
18452749	Comparative impact of 2 botulinum toxin injection techniques for elbow flexor hypertonia.	N	3
15191812	Tat-calpastatin fusion proteins transduce primary rat cortical neurons but do not inhibit cellular calpain activity.	Y	2
19336459	Motor cortex stimulation for the treatment of refractory peripheral neuropathic pain.	N	2
17608322	Pediatric CI therapy for stroke-induced hemiparesis in young children.	N	2

Table 10.10: The ten *least* commonly-selected articles under the Bilingual interface mode.

	En	Bi	Es
Recall	0.650 (0.312)	0.714 (0.353)	0.688 (0.336)
Precision	0.714 (0.170)	0.538 (0.239)	0.632 (0.196)
F-Measure	0.645 (0.238)	0.615 (0.272)	0.632 (0.246)

Table 10.11: Median classification accuracy scores, using the “Has TBI” computed relevance judgments. Compare to Table 9.25.

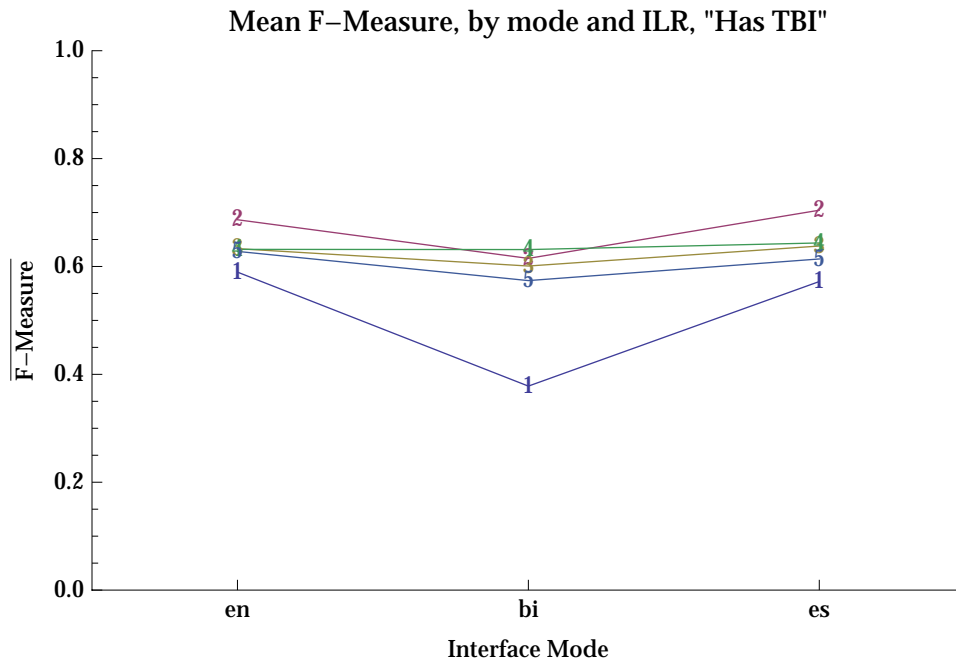


Figure 10.7: Mean F-measure, by ILR and interface mode, using the “Has TBI” computed relevance judgements. Compare to Figure 9.24.

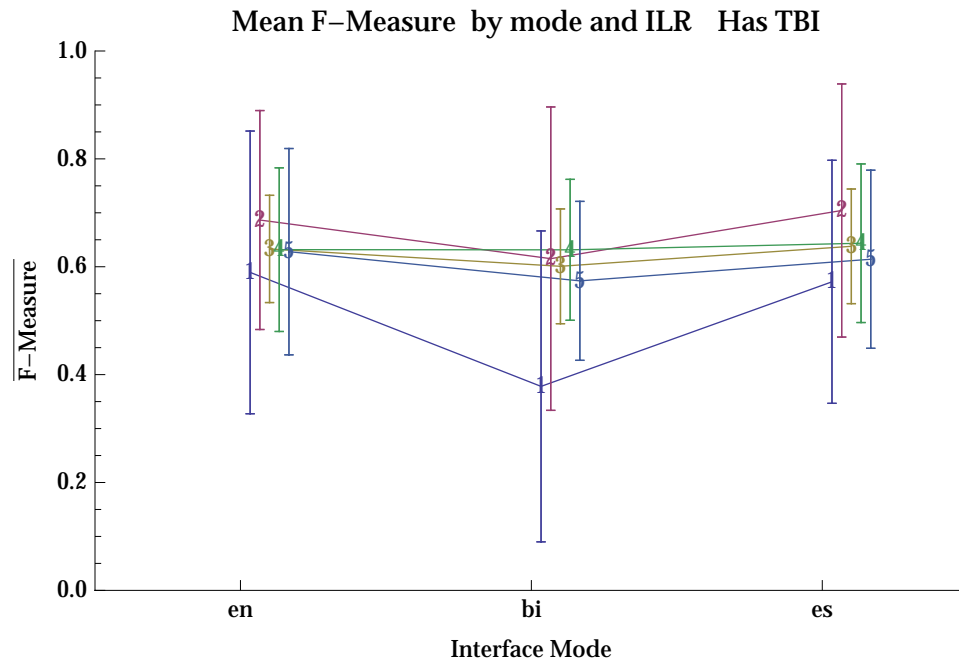


Figure 10.8: Mean F-measure, by ILR and interface mode, using the “Has TBI” computed relevance judgements and including error bars representing 95% confidence intervals. Compare to Figure 9.25.

10.3 Speed

There exist significant limitations regarding our data describing the amount of time subjects spent using the various interface modes while participating in the document selection task. First and foremost, as was discussed in Section 8.3, our experimental design did not enable us to fully attribute any observed differences in time spent between interface modes to the effect of the interface itself. Since we imposed no time constraints, each subject was free to spend as much or as little time as they liked with each interface; if they happened to spend more time under one interface than another, one possible interpretation of that fact would indeed be that they did so as a result of the difference in user interfaces. However, other explanations are possible; see the previously-referenced section’s discussion of speed/accuracy tradeoffs. As such, we must be very cautious in interpreting our speed data, as it is entirely possible that what may appear to be differences between interface modes

could in fact simply be artifact resulting from between-subject variation in how subjects responded to a task without time constraints.

A second important limitation to interpreting our speed data is that, as was described in the previous section, there exists a great deal of doubt regarding what our subjects thought that they were doing as they participated in the document selection task. As such, it is entirely possible that different groups of subjects might have interpreted our instructions differently from one another, making any between-subject comparisons essentially meaningless.

However, as was also discussed, while they might not have been carrying out the precise task that we had intended, whatever they *were* doing was far from random, and they spent a considerable amount of time doing it (see Table 9.27). Furthermore, there is no reason to suppose that an individual subject's conception of what their task was would change from interface mode to interface mode—in other words, whatever task they thought they were performing under the monolingual English interface mode, it seems likely that they would have been performing the same task under the Bilingual and monolingual Spanish interface modes. Since our metrics relating to speed were entirely based on *within-subject* changes from one interface mode to another, our findings may yet be useful.

As described in Section 9.4.2, our subjects tended to spend less time using the Monolingual Spanish interface—in which the results were entirely machine-translated—than the other two interfaces, by a median of nearly 28 seconds. Due to the aforementioned limitations, we cannot definitely interpret this finding to mean that our subjects were actually able to complete the document selection task most quickly under the Monolingual Spanish interface mode. However, we can safely say that our subjects spent less time using the Monolingual Spanish mode than the other two modes; while we cannot say for certain what the subjects were *doing* with that time, we can at least measure how much time they spent doing

it, and it appears as though, for whatever reason, our subjects did whatever they were doing more quickly under the Monolingual Spanish interface than under the other two interfaces.

These data are interesting for two reasons. First, recall that our initial hypothesis was that subjects would complete the task most quickly under the Monolingual Spanish interface; while our data did not support this hypothesis (due to the fact that we are unsure as to what task it was that subjects were attempting to complete), it is certainly suggestive that the mode we expected to prove fastest was indeed the one that subjects spent the least amount of time on. This remains an important area for future experiments to investigate.

The second reason these data are interesting is that, in addition to *objectively* measuring how long subjects spent selecting articles on each interface, we also asked subjects to tell us which of the three interface modes they *subjectively* felt fastest at using, in the form of the usability feedback question “I felt fastest at using the (blank) interface” (see Section 10.1 for further discussion of the usability feedback questions). We may therefore examine how accurate subjects’ perceptions of their own speeds were by comparing their responses to the subjective usability feedback questions with our objective measurements of how long they spent using each mode. While this method is far from perfect, it still yields interesting findings.

As shown in Table 10.12, many subjects who *felt* fastest when using the Monolingual English interface mode were actually fastest (or, at the very least, spent the least time using) at using the Monolingual Spanish interface mode. This effect is even more dramatic when we collapse the Bilingual and Monolingual Spanish modes together, and compare the “English-only” interface to interfaces containing “Some Spanish” (Table 10.13). Of the 18 subjects included in this table who reported that they felt the fastest when using the Monolingual English mode, fully 14 (just over 75%) actually spent less time on one of the Spanish-containing interfaces

		Perceived Fastest		
		English-only	Bilingual	Spanish-only
Actual Fastest	English-only	4	7	4
	Bilingual	4	3	4
	Spanish-only	10	7	8

Table 10.12: Subjects *perceived* fastest interface modes crosstabulated with their *actual* fastest interface mode.

		Perceived	
		English-only	Some Spanish
Actual	English-only	4	11
	Some Spanish	14	22

Table 10.13: Subjects *perceived* fastest interface modes crosstabulated with their *actual* fastest interface mode, collapsed into “Some Spanish” and “English-only.”

than on either of the other two interfaces. Conversely, only 33% of subjects who felt fastest at using one of the two Spanish-containing interfaces actually turned out to spend the least amount of time on the Monolingual English interface mode.

What does this mean? First of all, many subjects’ perceptions of which mode they were fastest at were inaccurate, but it was relatively more common for subjects who felt fastest at the Monolingual English mode to turn out to actually “be fastest” at one of the Spanish modes than for subjects who felt fastest when using a Spanish-containing interface to end up actually being faster at using the Monolingual English interface mode. One interpretation of this finding is that some users who might not *think* they would benefit from using a translated user interface would, in reality, benefit in some way.

We must be cautious here, as there are several possible explanations—perhaps subjects who felt fastest at the English-only interface were more likely to simply “breeze through” when using the Spanish interfaces, and thereby appear to be “faster” (when in reality they were not performing the document selection task

in an equivalent manner; see the discussion of speed-accuracy tradeoffs in Section 8.3). However, the fact that subjects tended, on average, to select *more* articles when using the Monolingual Spanish mode than the other two modes, and an equal number of articles when using the Bilingual mode, suggests that this explanation may not be correct; however, for these data to truly refute this explanation, we would need to be able to show that subjects performed the task with equivalent (or better) accuracy under one of the Spanish-containing modes, not just that they selected as many or more documents.

One other interesting observation from Table 10.12 concerns the bilingual interface mode. Very few of the subjects who felt fastest at using the bilingual mode actually were so, and the speed-related *perceptions* of those subjects who actually *were* fastest at the bilingual mode were evenly divided across the three interface modes. As with all of our other findings regarding the bilingual interface mode, this finding is somewhat perplexing, as the other two interface modes did not see this bifurcation. As in other aspects of our study (user preference, in particular), the bilingual interface mode has proven in this case to be something of a “wild card,” in that it did not follow the same pattern as did the monolingual interface modes. Whether this is an artifact of our data or an actual finding is a matter for future studies to determine.

In the end, we must conclude that, as was the case with classification accuracy, our speed-relating findings are intriguing, but are ultimately inconclusive. We believe that this area represents a very promising focus for future experimentation.

10.4 Limitations

This study had numerous limitations. First and foremost is that of sample size; while 59 subjects is a large amount by the standards of information retrieval user

studies, it is too small to allow us to draw any unambiguous conclusions about any of our hypotheses. Furthermore, our subject population was extremely diverse, professionally, geographically, and linguistically. As such, our population variance was very high for many of our metrics. This was primarily a result of our recruitment strategy, which, by necessity, was somewhat scattershot in nature. While this methodology was appropriate for a early-stage study such as this one, future studies would benefit from a more focused recruitment strategy and a less diverse population.

One area in which we would particularly like to improve our recruitment is among users with lower levels of English proficiency. Our data suggest that these users may be most likely to benefit from and prefer interfaces with translated content, but because we ended up with relatively few subjects in the lower ILR bands, we were unable to conclusively demonstrate that this is the case. The ILR instrument seemed to work reasonably well as an assessment tool, and may prove to be a helpful aid in determining which subjects to enroll.

The second major limitation to our study lies in the fact that we were essentially unable to address one of our primary hypotheses due to our inability to reliably interpret our document selection data. This means that we are unable to answer the question of how translated search interfaces affect users' ability to identify relevant articles. What caused this state of affairs? We may point at several contributing factors.

First and foremost, our document collection ended up being far noisier in reality than we had originally thought it to be, as was the semi-automatically-generated ground truth. While this situation may have been prevented by using more precise PubMed queries to generate our collection, it is quite possible that our Clinical Queries-based approach to collection development is fundamentally flawed. Future work may include attempting to improve this methodology.

A second component cause of our document selection difficulties lies in the task itself. From our experience with our secondary raters, we learned that our task, as we designed it, was much more ambiguous than our initial testing had led us to believe (and than we had intended it to be). Recall that this initial testing was done with subject experts accustomed to systematically reviewing collections of articles, and who were therefore familiar with the process of selecting or rejecting articles according to some specific set of inclusion criteria. Two of our three secondary raters had relatively limited experience with such practices, and indeed those two individuals independently described to us their ambiguous feelings about many of their relevance judgments. When we designed the task, we thought that the concept of “diagnosis” was distinct enough that it would be easy to identify; it seems as though we were incorrect in thinking so. Of course, it is possible that the problem lay not with the “diagnosis” decision rule itself but with the inconsistent nature of our document set. Perhaps, had the articles been less heterogeneous, our subjects and raters would have been less unsure about which were “diagnostic” in nature.

A third potential cause of our difficulties could lie in the instructions that we provided to our subjects.⁷ Obviously, at the time that the study was conducted, we felt that our instructions were fairly straightforward and simple; judging from the apparent behavior of our subjects,⁸ however, we can only conclude that our instructions were not sufficiently clear. Without being able to conduct follow-up interviews with our subjects, we cannot know for sure what their interpretations of the instructions were.

In future studies, we will attempt to address the issue of instruction clarity this in several ways. First, we will expand our instructions to include a TREC-

⁷ See Section 8.2 for the actual instructions themselves.

⁸ Specifically, the fact that many of the most frequently-selected articles were in no way, shape, or form about the diagnosis of traumatic brain injury.

style “narrative” describing the relevance criteria in greater detail and including exposition on which specific sorts of article might be relevant and which would not. We will also perform additional qualitative testing of our instructions with subjects drawn from the same recruitment pool as the rest of our subjects. This will help us identify any problems relating to wording or translation.

The second way we plan to address instruction clarity, along with improvements to the textual instructions themselves, will be to provide subjects with annotated examples of both relevant and non-relevant articles along with our instructions; in this study, our only example article was actually part of a screenshot demonstrating how to use the selection interface, and was not accompanied with any guidance about relevance or non-relevance. In addition to these examples, we will add a “training” step to our evaluation, in which subjects will be “walked through” the process of selecting articles, along with feedback about the accuracy of their selections. This should help ensure that our subjects are looking for the same thing that we think they should be.

By mitigating these three issues— the contents of the document collection, the unclear nature of the task, and the potentially inadequate instructions that we provided to our subjects— we believe that our underlying methodology (i.e., using a document selection task) may ultimately prove useful in investigating the effects of translated interfaces on user performance.

Another limitation to our study lies in the large number of subjects who “dropped out” of our study (i.e., stopped participating part-way through), as shown in Figure 9.1. We have no way of knowing precisely why these subjects chose to leave; however, it is certainly fair to note that our study was not a short one, and in fact took many of our subjects much longer than we had anticipated. Our ability to measure how long subjects took to complete the study are somewhat coarse-

grained,⁹ but we can report that our 59 usable subjects took a median of 48 minutes (std. dev.: 32 min.) to complete the study. This is longer than we had initially expected (and longer than our recruitment letter indicated our study would take), and we would not at all be surprised to learn that many of our drop-out subjects did so because the study was taking too much time.

To counteract this, future studies will use much more streamlined protocols than did this one. The survey instruments we used for this study contained many questions that were exploratory in nature, and while they provided useful background information about our subjects, they did not ultimately prove helpful in addressing our central hypotheses. Fewer and shorter survey instruments could help cut down the time spent by subjects dramatically. Furthermore, our protocol called for a total of five document selection tasks: a warmup task, the three experimental tasks, and then a followup task. In the end, the followup task did not prove informative, and could easily be dropped from the protocol. Some sort of warmup task would still be necessary in a revised protocol, but we feel that it could be merged with the “training” step outlined above. While we can not be sure that a shorter protocol would have resulted in fewer dropouts, it certainly would not hurt, and even if it resulted in no additional subjects, it would make participating in our study into less of a burden for those subjects who did complete the protocol.

An additional limitation to our study overall lies in the issues surrounding our speed metrics (discussed at length in sections 8.3 and 10.3). Put simply, because

⁹ We are able to tell when they first arrived at the site, and when they ultimately completed the last step; when we initially designed the FREDO evaluation-management system, we anticipated that these data points would be sufficient. However, using these data, many of our subjects appeared to take an unrealistically long time to complete the study (in some cases, more than 12 hours). Upon investigating these cases more closely, we invariably discovered that these subjects stopped in the middle of one step, and then resumed the step several hours later. In other words, many— in fact, most— of our subjects appear to have participated in fits and starts over the course of an afternoon, evening, or entire day. Of course, given that our subjects were busy professionals, it is unsurprising to discover that many of them did not have extended blocks of uninterrupted time to devote to our study. Future versions of the FREDO system will include an improved approach to timing subject participation.

we did not impose any time limits on our subjects, and indeed told them to take as long as they felt was appropriate, it is difficult to draw conclusions from our speed data. One simple way to address this issue in future evaluations would be to keep our task the same, but impose some sort of time limit. Then, rather than using elapsed time as the outcome measure, we would instead use some variation on the number of articles reviewed by the subject during their allotted time. This approach has its appeal, although we would need to do considerable experimentation to calibrate both the time limit as well as the size of the document collection.

A final limitation involves the way that our selection interface was implemented. In order to indicate that they considered a particular article to be relevant, subjects had to perform an action: they had to actively click on that article's link. For the purposes of this study, we considered any articles that the subject left blank (i.e., those articles that the subject chose not to actively select) to be not-relevant. When we were designing the selection interface, we did not think of this lack of action on the user's part in passive terms; we assumed that the subjects would give each article equal consideration, and that their choice (to mark any given article, or to leave it un-marked) would be an active and conscious one. This may not have been a safe assumption. If a subject left an article un-selected, that could indeed mean that they did not feel it to be relevant... but it could also mean that they did not read the article, and instead simply skipped on to the next one in the list.

As such, it is possible that our count of the "not relevant" judgments made by our subjects could be inflated, biasing some of our subjects in the direction of an increased false negative rate. Also, there were a small number of subjects who made no document selections at all when using one or more of the interfaces; we have no way of definitively knowing whether they honestly felt that none of the articles were relevant, or whether they simply skipped to the bottom of the screen and moved on. To mitigate this problem, future iterations of the interface will be

redesigned such both “relevant” and “not relevant” judgments will require active engagement on the part of the users. This approach is not without issues, however. Some number of subjects will inevitably fail to record either judgment, and will leave some number of articles as un-marked. Any experimental protocols using this interface will need to account for these cases.

We believe that, while significant, these limitations represent major opportunities for future refinement of our protocol and system. Some of them are obvious in hindsight; others would have been difficult to uncover were it not for this study and the experience we gained by fielding our experimental protocol and system with real-world subjects. We feel that we have laid a solid foundation, and have a clear path on which to move forward towards the next iteration of this protocol. We are hopeful that future studies using our various systems and instruments will be able to address the questions that this one was unable to conclusively answer.

Chapter 11

Conclusion & Future Work

In spite of the limitations described in Section 10.4, this dissertation achieved many of its aims, and represents a novel contribution to the field of biomedical informatics. To demonstrate the feasibility of fully integrating a commercial machine translation system with a powerful bibliographic search system, we built and deployed the BuscTrad bilingual literature search system,¹ as well as a novel tool designed to support and manage internet-mediated user studies. Using these tools, we conducted a small-scale user study involving a wide variety of Latin American clinical professionals from eleven different countries. In the course of this study, we discovered that, given the choice between translated and non-translated search results, many non-native English speakers preferred translated results, even in the face of the significant issues described in Section 9.3. Furthermore, we found that a sizable portion of our subjects expressed a preference for bilingual presentation of search results over monolingual result presentation.

More importantly, we discovered that users' preferences regarding translated content are heavily dependent on their individual levels of English reading proficiency. Subjects with weaker English reading abilities were much more likely to ex-

¹Currently available for public use at: <http://skynet.ohsu.edu/busctrad>

press a preference for one of the translated interface modes than were subjects with stronger English abilities. However, this was definitely not an “all-or-nothing” pattern: many of our subjects with higher ILR scores expressed preferences for one of the translated interface modes. This finding is relevant for future developers of multilingual search systems, as is the simple fact that there was so much diversity of opinion among our subjects. This suggests that users appreciate having a variety of user interface options when reviewing foreign-language search results.

In addition to developing and fielding a novel literature search interface, this study also demonstrated that the Interagency Language Roundtable’s English reading self-assessment tool can be easily used as part of a user study, and that its results appear to have some experimental validity. Furthermore, we demonstrated that the tool can be used effectively even when translated into a different language than that in which it was originally developed. Other researchers needing to quickly and cheaply estimate their subjects’ levels of English reading expertise may use this instrument with some degree of confidence, knowing that it has been field-tested. Furthermore, the ILR produces several other self-assessment tools (for other language competencies, such as speaking and listening), and the fact that the reading assessment instrument worked reasonably well in translation lends credibility to the idea that these other instruments might similarly function well in translation.

In the end, we were unable to determine conclusively whether or not our translated interfaces affected our subjects’ ability to accurately identify relevant articles; what evidence we have *suggests* that, at the very least, it does not *hurt* their ability to do so, but we really cannot say one way or the other with any degree of confidence. As discussed in chapter 4, ours is not the first such study to suffer from inconclusive or confusing results. However, in the process of investigating what went wrong with this portion of our study, we identified many issues, both

with our experimental design and with our study materials, and learned numerous lessons about how to effectively conduct this sort of research. The challenges we experienced underscore the difficulties and complexities inherent in evaluating interactive information retrieval systems, particularly across linguistic and geographical divides. Ultimately, it is only by performing experiments with actual users that many of these issues may be discovered and addressed, and the field moved forward.

We have numerous plans for future work along these lines. First and foremost of these is that of a similar study to this one, albeit with the various modifications discussed in Section 10.4. In addition to this study's main findings, in the course of carrying out this study we obtained useful demographic and linguistic data about our intended user population, which we will be able to use to help design and power the next round of studies. Our Latin American partners have enthusiastically offered to help obtain subjects for future studies, and now that we know more about the various populations that we will be working with, we will be better able to estimate how many subjects we will need, as well as which specific groups to target for recruitment. We are hopeful that including more focused and homogeneous groups of subjects will help us avoid some of the issues that plagued the present study.

Once we have refined our experimental protocol and survey instruments, and have addressed the various issues as previously discussed, we hope to branch out beyond Latin America, and explore whether our findings are consistent across languages. Do Russian-speaking clinicians respond the same way to translated content as Spanish-speaking clinicians? What about Arabic-speaking clinicians? Do speakers of some languages benefit more (or less) from translated content?

In addition to experimenting with new languages, we would also like to experiment with different translation techniques. In this study, for a variety of reasons,

we used translations provided by Google Translate. As our subjects were only too happy to point out, these translations were far from perfect. We are curious as to whether our results would be different if we used a different machine translation engine, one that might provide better (or worse!) translations. Also, for this study, we used a general-purpose machine translation system; we would like to investigate whether or not a system specifically tuned to translate medical text might perform better. Furthermore, we are curious as to whether our protocol might ultimately be useful for obtaining user-oriented evaluations of the translations provided by different machine translation systems— addressing the question of whether system “A” is measurably “better” than system “B,” from a user’s perspective. The quality of machine-translated text is very difficult to quantitatively evaluate in a way that is meaningful to human users, and having a reliable framework for conducting user studies along these lines has the potential to be very useful.

In conclusion, this study achieved some degree of success while at the same time highlighting a number of challenges and directions for future work. We addressed many of our goals; others remained elusive. In the course of our experiment, we encountered numerous obstacles and challenges, many of which were not specific to our study but rather could easily be repeated by other information retrieval researchers, particularly those who venture into the arena of cross-cultural research. It is therefore our hope that this account of the various and sundry crevasses, swamps, deserts, and sink-holes that we encountered during this study may prove useful to future investigators, so that their studies might follow a straighter, smoother, drier, and altogether less hazardous road than did our own.

Appendix A

Final FREDO evaluation plan

Listing A.1: The final evaluation plan for the study protocol described in chapter 8.

```
1
2 eval_plan('final') do |e|
3
4   # intro
5   e.page(10, :es)
6
7   # phase 1 (demographic survey)
8   e.survey(75, :es)
9
10  # phase 2 (language survey)
11  e.survey(68, :es)
12
13  # instructions
14  e.page(9, :es)
15
16  doc_selection_base_url = "http://skynet.ohsu.edu/busctrad/canned"
17
18  # warmup selection task
```

```
19   e.external(doc_selection_base_url,
20     {:article_set_id => 39, :lang => 'en'}
21   )
22
23   # latin square block
24   e.latin do |1|
25
26     # english
27     l.external(doc_selection_base_url,
28       {:article_set_id => 40, :lang => 'en'}
29     )
30
31     # spanish
32     l.external(doc_selection_base_url,
33       {:article_set_id => 41, :lang => 'es'}
34     )
35
36     # bilingual ("side-by-side")
37     l.external(doc_selection_base_url,
38       {:article_set_id => 42, :lang => 'sbs'}
39     )
40
41   end
42
43   # follow-up step
44   e.external(doc_selection_base_url,
45     {:article_set_id => 43, :lang => 'en'}
46   )
47
48   # follow-up (feedback) survey
49   e.survey(70, :es)
50
51   # final "thanks! page"
```

52 e.page(3, :es)

53

54 **end**

Appendix B

Survey Instruments: Demographics, English

0. A. Personal Information

1. How old are you?

2. What is your gender?

(Choose one)

Female

Male

Not Specified

3. What country were you born in?

4. What country do you currently work in?

5. On a scale of 1-5, how would you rate your computer skills?

(Choose one)

1 - Unable to use a computer

2

3

4

5 - Expert computer user

6. Please check the box next to any of the following types of software that you regularly use.

(Multiple choice)

Database software

- Statistical software
- Word processor software
- Spreadsheet software
- Spreadsheet software
- E-Mail software
- Messenger/Gmail Chat
- Other

7. Other

8. How often do you use a computer for non-work reasons?

(Choose one)

- Every day
- Between 3 and 6 days per week
- 1 or 2 days per week
- One or two days per month
- I don't use computers for non-work reasons.

9. In the last week, how many hours have you spent using a computer outside of work?

(Choose one)

- Less than one hour
- Between 1 and 5 hours
- Between 5 and 10 hours
- Between 10 and 20 hours
- More than 20 hours

10. How often do you use an Internet search engine (Google, Yahoo, etc.) to look for non-work information?

(Choose one)

Every day

Between 3 and 6 days per week

1 or 2 days per week

One or two days per month

I don't use search engines for non-work Internet searches.

11. What search engine do you use most often for non-work Internet searches?

12. On a scale of 1-5, how would you rate your ability to use Internet search engines to find information?

(Choose one)

1 - No ability

2

3

4

5 - Expert searcher

13. Do you use social networking sites (Facebook, Orkut, MySpace, Twitter, etc.)?

(Choose one)

Yes

No

14. What network do you use most often?

(Choose one)

Facebook

MySpace

Orkut

LiveJournal

Twitter

Other

15. Other

16. B. Professional Information

17. Are you a...

(Choose one)

Medical Student

Resident

Nurse

Attending Physician

Other

18. Other

19. How many years have you been in your current role? (Student, Nurse, etc.)

20. What is your medical specialty (if you have one)?

21. On a scale of 1-5, how would you rate your level of knowledge about traumatic brain injury *diagnosis*?

(Choose one)

1 - I know nothing about diagnosing traumatic brain injury

2

3

4

5 - I am an expert at diagnosing traumatic brain injury

22. On a scale of 1-5, how would you rate your level of knowledge about traumatic brain injury *treatment*?

(Choose one)

1 - I know nothing about treating traumatic brain injury

2

3

4

5 - I am an expert at treating traumatic brain injury

23. How often do you use a computer for work-related reasons?

(Choose one)

Every day

Between 3 and 6 days per week

- 1 or 2 days per week
- One or two days per month
- I don't use computers for work-related reasons.

24. In the last week, how many hours have you spent using a computer at work?

(Choose one)

- Less than one hour
- Between 1 and 5 hours
- Between 5 and 10 hours
- Between 10 and 20 hours
- More than 20 hours

25. How often do you use an Internet search engine (Google, Yahoo, etc.) to look for work-related information?

(Choose one)

- Every day
- Between 3 and 6 days per week
- 1 or 2 days per week
- One or two days per month
- I don't use search engines for work-related Internet searches.

26. What search engine do you use most often for work-related Internet searches?

27. How often do you look up medical information using an electronic resource of some kind (medical textbook, clinical guideline, published article, etc.)?

(Choose one)

Every day

Between 3 and 6 days per week

1 or 2 days per week

One or two days per month

I don't use electronic resources to look up medical information.

28. Which of the following resources have you used in the last month?

(Multiple choice)

Electronic Textbook

MEDLINE (via PubMed)

MEDLINE (via some other source)

Embase

WHO-HINARI

US CDC MMWR

Other US CDC Reference

Cochrane Library, English

Cochrane Library, Spanish

Other

29. Other

Appendix C

Survey Instruments: Demographics, Spanish

0. A. Información Personal

1. ¿Cuántos años tiene?

2. ¿Cual es su sexo?

(Elege uno)

Femenino

Masculino

No sabe/no contesta

3. ¿En qué país nació?

4. ¿En qué país trabaja actualmente?

5. En una escala de 1-5, ¿cómo calificaría usted sus conocimientos de computación?

(Elege uno)

1 - Incapaz de utilizar una computadora

2

3

4

5 - Experto en el uso de computadoras

6. Señale los programas que usted utiliza habitualmente.

(Opción múltiple)

Base de datos

- Programas estadísticos
- Procesadores de texto
- Planillas de calculo
- Planillas de calculo
- Correo electrónico
- Messenger /Gmail Chat
- Otro

7. Otro

8. ¿Con qué frecuencia utiliza la computadora para tareas no relacionadas con el trabajo?

(Elege uno)

- Todos los días
- Entre 3 y 6 días por semana
- Entre 1 y 2 días por semana
- Entre 1 y 2 días por mes
- No uso computadoras para tareas no relacionadas con el trabajo.

9. En la última semana, ¿cuánto tiempo usó la Computadora para tareas no relacionadas con el trabajo?

(Elege uno)

- Menos de 1 hora
- Entre 1 y 5 horas
- Entre 5 y 10 horas
- Entre 10 y 20 horas
- Mas de 20 horas

10. ¿Con qué frecuencia utiliza un buscador de Internet (Google, Yahoo, etc) para buscar información no relacionada al trabajo?

(Elege uno)

Todos los días

Entre 3 y 6 días por semana

Entre 1 y 2 días por semana

Entre 1 y 2 días por mes

No uso buscadores de Internet para buscar información no relacionada al trabajo.

11. ¿Qué buscador de internet es el que mas utiliza para buscar informacion no relacionada al trabajo ?

12. En una escala de 1-5, ¿cómo calificaría usted su capacidad para utilizar los buscadores de Internet para encontrar información?

(Elege uno)

1 - Ninguna capacidad

2

3

4

5 - Experto/Maxima capacidad

13. ¿Utiliza un sitio de redes sociales como Facebook, Orkut, MySpace, Twitter, etc.?

(Elege uno)

Si

No

14. ¿Cual es la red social que utiliza con más frecuencia?

(Elege uno)

Facebook

MySpace

Orkut

LiveJournal

Twitter

Other

15. Otro

16. B. Información Profesional

17. ¿Es usted...

(Elege uno)

Estudiante de Medicina

Médico Residente

Enfermera

Médico

Other

18. Otro

19. ¿Cuántos años ha sido usted un (estudiante, enfermera, etc)?

20. ¿Cuál es su especialidad médica? (si tiene uno)

21. Con una escala de 1-5, ¿como calificaría su nivel de conocimiento sobre el *diagnóstico* de lesiones cerebrales traumáticas?

(Elege uno)

1 - No sé nada sobre el diagnóstico de lesiones cerebrales traumáticas.

2

3

4

5 - Soy un experto en el diagnóstico de lesiones cerebrales traumáticas.

22. Con una escala de 1-5, ¿como calificaría su nivel de conocimiento sobre *tratamientos* de lesiones cerebrales traumáticas?

(Elege uno)

1 - No sé nada acerca de los tratamiento de lesiones cerebrales traumáticas.

2

3

4

5 - Soy un experto en el tratamiento de lesiones cerebrales traumáticas.

23. ¿Con qué frecuencia utiliza la computadora para razones relacionadas al trabajo?

(Elege uno)

Todos los días

- Entre 3 y 6 días por semana
- Entre 1 y 2 días por semana
- Entre 1 y 2 días por mes
- No uso computadoras para el trabajo.

24. En la última semana, ¿cuánto tiempo usó la Computadora para tareas para el trabajo?

(Elege uno)

- Menos de 1 hora
- Entre 1 y 5 horas
- Entre 5 y 10 horas
- Entre 10 y 20 horas
- Mas de 20 horas

25. ¿Con qué frecuencia utiliza un buscador de Internet (Google, Yahoo, etc) para buscar información relacionada al trabajo?

(Elege uno)

- Todos los días
- Entre 3 y 6 días por semana
- Entre 1 y 2 días por semana
- Entre 1 y 2 días por mes
- No uso buscadores de Internet para buscar información relacionada al trabajo.

26. ¿Qué buscador de internet es el que utiliza con mas frecuencia para buscar informacion relacionada a su trabajo?

27. ¿Con qué frecuencia utiliza un recurso electrónico (libros de texto médicos, guía clínica, artículos publicados, etc.) para buscar información médica?

(Elege uno)

Todos los días

Entre 3 y 6 días por semana

Entre 1 y 2 días por semana

Entre 1 y 2 días por mes

No uso recursos electrónicos para buscar información médica.

28. ¿Cuál de los siguientes recursos ha utilizado en el último mes?

(Opción múltiple)

Libros de textos electrónicos

MEDLINE (a través de PubMed)

MEDLINE (a través de algún otro sitio)

Embase

OMS-HINARI

EEUU CDC MMWR

Otros recursos de los CDC

Cochrane Library, Inglés

Cochrane Library, Español

Otro

29. Otro

Appendix D

ILR Self-Assessment Scale, English

The ILR Reading Self-Assessment tool can be found at: <http://www.govtilr.org/Skills/readingassessment.pdf>. Note that the contents of the scale were translated into Spanish for use in this evaluation (see Appendix E).



SELF-ASSESSMENT OF READING PROFICIENCY

The following Self-Assessment of foreign language Reading Ability is intended to serve as a guide for people who have not taken a U.S. Government-sponsored reading test but would like to have a rough estimate of their proficiency. The self-assessment questionnaire will produce an estimate of your current foreign language reading ability but is in no way intended to be a replacement for the existing ILR Skill Level Descriptions.

Important: The term *read* as used in this self-assessment always means “*read and understand the meaning.*” It does not refer in any way to the ability to read aloud without comprehension. The term *text* refers to any example of language presented in the writing system of the language, including advertisements, weather reports, news articles, letters, lengthy essays, and literary works, among others.

For all texts at a level, it is not necessary to know all the words or understand all the details of the texts listed for that level, but it is necessary to perform the functional tasks described for the level at the indicated level of accuracy.

To estimate your level of proficiency, start at the lowest level (R-0+) and respond to each statement. For each statement, respond “yes” or “no.” If a statement describes your ability only some of the time, or only in some contexts, you should answer “no.” If you answer “yes” to every statement in the level, your ability is probably at least at that level. Move on to the descriptions at the next level. If you answer “no” to one or more statements, then you are likely not at that level.

If you answer “yes” to all the statements at one level, and have a majority of “yes” answers at the next higher level, then you may be at a “plus” level. For example, if you answer “yes” to all the statements at Level 1, but have a mixture of responses at Level 2 (almost all “yes” answers), your self-assessed ability may be at Level 1+.

Note to the user: This self-assessment instrument is posted by the ILR in provisional form for personal use by any interested individual. The final version will be posted after one year. Please send any comments or suggestions for improving the form by no later than February 15, 2010, to Dr. Frederick H. Jackson (fjackson@nflc.org).

SELF-ASSESSMENT OF READING PROFICIENCY		Yes	No
R-0+	As appropriate for the language, I can recognize and identify all the letters in the printed version of an alphabetic writing system (in languages like English, Spanish, Finnish, Russian, Greek, Vietnamese) or the elements of a syllable-based writing system (such as in Japanese kana, Korean hangul, Hebrew, Arabic, Amharic, Thai, or Hindi) or some commonly occurring characters in a character system (Chinese, Japanese kanji, Korean hanja.)		
R-0+	I can read some isolated words and phrases, such as numbers and commonplace names, that I see on signs, menus, and storefronts, and in simple everyday material such as advertisements and timetables.		
R-1	I can understand the purpose and main meaning of very short, simple texts, such as in printed personal notes, business advertisements, public announcements, maps, etc.		
R-1	I can understand simple instructions, such as in very straightforward street directions.		
R-1	I can understand very short simple written descriptions of some familiar persons, places, and things, like those found in many tourist pamphlets.		
R-2	I can understand texts that consist mainly of straightforward factual language, such as short news reports of events, biographical information, descriptions, or simple technical material.		
R-2	I can understand the main idea and some details of clearly organized short straightforward texts about places, people, and events that I am familiar with.		
R-2	I can understand very straightforward reports about current and past events.		
R-2	I can understand simple typed correspondence in familiar contexts, including descriptions of events, feelings, wishes and future plans.		
R-2	I can usually understand the main ideas of authentic prose on topics I am familiar with, either because they pertain to my work experience or to topics I am interested in.		
R-3	I can usually read and understand all of the material in a major daily newspaper published in a city or country with which I am familiar.		
R-3	In reading a newspaper or magazine that contains editorial or opinion content, I can “read between the lines” and understand meanings that are not directly stated.		
R-3	I can understand the author’s intent and follow the line of reasoning in texts that include hypothesis, persuasion, supported opinion or argument for a position (e.g., editorials,		

	debates, and op-ed pieces) with little or no use of a dictionary.		
R-3	I can understand contemporary expository essays and recent literary prose with little or no use of a dictionary,.		
R-3	I can understand the main ideas and important details of almost all material written within my particular professional field or area of primary interest (e.g., reports, analyses, letters, arguments, etc.).		
R-4	I am able to read fluently and accurately all styles and forms of the language pertinent to professional needs or personal interest without reference to a dictionary,.		
R-4	I can understand long and complex analyses, factual reports, and literary texts.		
R-4	I can understand both the meaning and the intent of most uses of idioms, cultural references, word play, sarcasm, and irony in even highly abstract and culturally “loaded” texts.		
R-4	I can understand language that has been especially adjusted for different situations, audiences or purposes, such as a political essay, humorous anecdote or joke, sermon, or inflammatory broadside, and I can appreciate distinctions in style.		
R-4	I can read virtually all forms of the written language, including abstract, linguistically complex texts such as specialized articles, essays and literary works, including prose works from earlier periods recognized as masterpieces.		
R-4	I can read reasonably legible handwriting without difficulty		

Appendix E

ILR Self-Assessment Scale, Spanish

1. Puedo reconocer e indentificar todas las letras del alfabeto.
2. Puedo leer algunas palabras y frases aisladas, como los números y nombres comunes, que veo en las señales, los menús, y tiendas, y en materiales sencillos y cotidianos como anuncios y publicidades.
3. Puedo entender el propósito y el significado principal de textos muy breves y simples, como notas personales escritas, anuncios comerciales, anuncios públicos, mapas, etc.
4. Puedo entender instrucciones sencillas, por ejemplo como llegar a un determinado lugar.
5. Puedo entender textos, cortos y simples sobre personas, lugares, y cosas conocidos como los que se encuentran en folletos turísticos.
6. Puedo entender textos fácticos y sencillos, tales como noticias breves del diario , datos biográficos, descripciones o material técnico simple.
7. Puedo entender la idea principal y algunos detalles de textos breves ,sencillos claramente organizados sobre personas, lugares y eventos de los cuales que

estoy familiarizado.

8. Puedo entender los informes simples y claros sobre eventos actuales y pasados.
9. Puedo entender correspondencias sencillas, escritas a máquina en contextos familiares, incluyendo la descripción de eventos, sentimientos, deseos y planes futuros.
10. Generalmente puedo entender las ideas principales de la prosa auténtica sobre temas que conozco, ya sea porque pertenecen a mi experiencia laboral o temas que me interesan.
11. Por lo general puedo leer y entender todo el material de un periódico de mayor circulación en la ciudad o país con el que estoy familiarizado.
12. En la lectura de un periódico o revista que contiene contenido editorial o de opinión, puedo leer “el entre las líneas” y comprender lo que no está directamente declarado.
13. Puedo entender la intención del autor y seguir la línea de razonamiento en los textos que incluyen hipótesis, persuasión, opinión justificada (por ejemplo, editoriales, debates y otros artículos de opinión), con poco o ningún uso de un diccionario.
14. Puedo entender ensayos contemporáneos y la prosa literaria reciente con poco o ningún uso de un diccionario.
15. Puedo entender las ideas principales y los detalles importantes de casi todos los materiales escritos dentro de mi área profesional o área de interés principal (por ejemplo, informes, análisis, cartas, argumentos, etc.).

16. Capaz de leer con fluidez y precisión todos los estilos y formas del lenguaje pertinentes a las necesidades profesionales o de interés personal sin referencia a un diccionario.
17. Comprendo análisis largos y complejos, informes fácticos, y textos literarios.
18. Puedo entender el significado y la intención de la mayoría de los usos de los modismos, las referencias culturales, juegos de palabras, el sarcasmo y la ironía, incluso en textos muy abstractos y culturalmente “cargados”.
19. Puedo entender el lenguaje que ha sido especialmente adaptado para diferentes situaciones, audiencias u objetivos, tales como un ensayo político, una anécdota humorística, broma, sermón, o críticas, y puedo apreciar las diferencias de estilo.
20. Soy prácticamente capaz de leer todas las formas de la lengua escrita, incluyendo artículos lingüísticamente complejos como artículos especializados, ensayos y obras literarias, incluyendo las obras en prosa de períodos anteriores reconocidas como obras maestras.
21. Puedo leer la letra razonablemente legible sin dificultad.

Appendix F

ILR Instrument Scoring Function

This function, `calcIlrLev[]`, computes ILR band scores. It takes as its argument a list whose first element is a subject identifier and whose remaining elements represent that subject's responses to the various statements in the ILR instrument, in the order in which they appear in the original instrument. If the subject "agreed" with the statement, it is encoded as a "1"; otherwise, it is encoded as a "2". Statements with which the subject neither agreed or disagreed were encoded as empty strings.

The function returns a list whose first element is the subject identifier, and whose second element is a list containing: a number (1–5) representing the highest band with whose statements the subject completely agreed, the textual identifier used by the ILR instrument to describe that band ("r0," "r1," etc.), and a string indicating whether or not the subject scored into a "Plus"-level.

```
1 ilrKey = {1 -> 'Agree', 2 -> 'Disagree'};
2 ilrLevels = StringSplit[#[[1]], '_' &
3     /@ {'r0_1', 'r0_2', 'r1_1', 'r1_2', \
4     'r1_3', 'r2_1', 'r2_2', 'r2_3', 'r2_4', 'r2_5', \
5     'r3_1', 'r3_2', 'r3_3', 'r3_4', 'r3_5', 'r4_1', \
6     'r4_2', 'r4_3', 'r4_4', 'r4_5', 'r4_6'}];
```

```

7
8 idxToLevel = MapIndexed[
9   #2[[1]] -> #1[[1]] &, ilrLevels ];
10
11 levelToIdx = # -> Flatten[ Position[
12   ilrLevels ,
13   {#, _}
14   ]] & /@ {''r0'', ''r1'', ''r2'', ''r3'', ''r4''};
15
16 Clear[ calcIlrLev ];
17
18 calcIlrLev [ subj_ ] :=
19   Module [{ levelCnt , levs , groupedLevCount , cutoffs ,
20     highestCompleteAgree , plusLevel },
21     levelCnt = Tally[
22       MapIndexed[
23         { (#2[[1]] /. idxToLevel), #1 /. ilrKey } &
24         , subj[[2 ;;]] ]
25       , #1[[1]] === #2[[1]] && #1[[2]] === #2[[2]] &
26       ];
27     levs = { ''r0'', ''r1'', ''r2'', ''r3'', ''r4'' };
28     groupedLevCount = Cases[ levelCnt , { {#, _}, cnt_ } ] & /@ levs ;
29     highestCompleteAgree = Max[
30       Position[
31         groupedLevCount ,
32         {
33           { {_, ''Agree''}, _ }
34         }
35       ]
36     ];
37     cutoffs = (Max[# /. levelToIdx] & /@ { ''r0'', ''r1'', ''r2'',
38       ''r3'', ''r4'' });
39     plusLevel = If[

```

```

40     levs [[highestCompleteAgree]] == 'r4', (* no r4 plus... *)
41     '...',
42     (* figure out next highest level; if majority is 'agree', '+' *)
43
44     Module [{tmpCount},
45     tmpCount =
46     Sort [Tally [
47         subj [[ cutoffs [[highestCompleteAgree] + 1 ];
48             cutoffs [[highestCompleteAgree + 1]]]
49     ]];
50     If [
51     tmpCount[[1, 2]]/Total[tmpCount[[All, 2]]] >= (1/2),
52     'Plus', '...'
53     ]
54     ]
55     ];
56     Return [{highestCompleteAgree, levs [[highestCompleteAgree]],
57     plusLevel}]
58     ];

```

Appendix G

Survey Instruments: Language, English

0. These questions are about your experiences with spoken and written languages, particularly English. Your answers in this section are very important and helpful, but if there are questions you do not wish to answer you may skip them.

1. Which languages do you regularly use outside of work?

(Multiple choice)

Spanish

Portuguese

English

French

Italian

Quechua

Guaraní

Aymara

Other

2. Other

3. Which languages do you regularly use at work?

(Multiple choice)

Spanish

Portuguese

English

French

Italian

Quechua

Guaraní

Aymara

Other

4. Other

5. The next several questions are about what you think of your own level of English proficiency. Please answer them using a scale of 1 to 5, with 1 being 'no proficiency' and 5 being 'expert'. Choose your answer from the menu next to each question.

6. When you encounter the English language in non-professional contexts, how would you rate your level of proficiency at:

7. reading English?

(Choose one)

1 - Cannot read any English

2

3

4

5 - Perfect English reading ability

8. writing English?

(Choose one)

1 - Cannot write any English

2

3

4

5 - Perfect English writing ability

9. speaking in English?

(Choose one)

1 - Cannot speak any English

2

3

4

5 - Perfect English speaking ability

10. listening to spoken English?

(Choose one)

1 - Cannot understand any spoken English

2

3

4

5 - Perfect English listening comprehension

11. When you encounter the English language in professional contexts:

12. Reading English text?

(Choose one)

1 - Cannot read any English

2

3

4

5 - Perfect English reading ability

13. writing English?

(Choose one)

1 - Cannot write any English

2

3

4

5 - Perfect English writing ability

14. speaking in English?

(Choose one)

1 - Cannot speak any English

2

3

4

5 - Perfect English speaking ability

15. *listening* to spoken English?

(Choose one)

1 - Cannot understand any spoken English

2

3

4

5 - Perfect English listening comprehension

16. As compared to your colleagues, do you think your level of overall English proficiency is

(Choose one)

Higher than average

About average

Lower than average

17. Check either "agree" or "disagree" next to the following statements about your ability to read English-language text.

18. I can recognize and identify all the letters in printed English text.

(Choose one)

Agree

Disagree

19. I can read some isolated words and phrases, such as numbers and common place names, that I see on signs, menus, and store fronts, and in simple everyday material such as advertisements and timetables.

(Choose one)

Agree

Disagree

20. I can understand the purpose and main meaning of very short, simple texts, such as in printed personal notes, business advertisements, public announcements, maps, etc.

(Choose one)

Agree

Disagree

21. I can understand simple instructions, such as in very straightforward street directions.

(Choose one)

Agree

Disagree

22. I can understand very short simple written descriptions of some familiar persons, places, and things, like those found in many tourist pamphlets.

(Choose one)

Agree

Disagree

23. I can understand texts that consist mainly of straight forward factual language, such as short news reports of events, biographical information, descriptions, or simple technical material.

(Choose one)

Agree

Disagree

24. I can understand the main idea and some details of clearly-organized short straight forward texts about places, people, and events that I am familiar with.

(Choose one)

Agree

Disagree

25. I can understand very straightforward reports about current and past events.

(Choose one)

Agree

Disagree

26. I can understand simple typed correspondence in familiar contexts, including descriptions of events, feelings, wishes and future plans.

(Choose one)

Agree

Disagree

27. I can usually understand the main ideas of authentic prose on topics I am familiar with, either because they pertain to my work experience or to topics I am interested in.

(Choose one)

Agree

Disagree

28. I can usually read and understand all of the material in a major daily newspaper published in a city or country with which I am familiar.

(Choose one)

Agree

Disagree

29. In reading a newspaper or magazine that contains editorial or opinion content, I can "read between the lines" and understand meanings that are not directly stated.

(Choose one)

Agree

Disagree

30. I can understand the author's intent and follow the line of reasoning in texts that include hypothesis, persuasion, supported opinion or argument for a position (e.g., editorials, debates, and op-ed pieces) with little or no use of a dictionary.

(Choose one)

Agree

Disagree

31. I can understand contemporary expository essays and recent literary prose with little or no use of a dictionary.

(Choose one)

Agree

Disagree

32. I can understand the main ideas and important details of almost all material written within my particular professional field or area of primary interest (e.g., reports, analyses, letters, arguments, etc.).

(Choose one)

Agree

Disagree

33. I am able to read fluently and accurately all styles and forms of the language pertinent to professional needs or personal interest without reference to a dictionary.

(Choose one)

Agree

Disagree

34. I can understand long and complex analyses, factual reports, and literary texts.

(Choose one)

Agree

Disagree

35. I can understand both the meaning and the intent of most uses of idioms, cultural references, word play, sarcasm, and irony in even highly abstract and culturally "loaded" texts.

(Choose one)

Agree

Disagree

36. I can understand language that has been especially adjusted for different situations, audiences or purposes, such as a political essay, humorous anecdote or joke, sermon, or inflammatory broadside, and I can appreciate distinctions in style.

(Choose one)

Agree

Disagree

37. I can read virtually all forms of the written language, including abstract, linguistically complex texts such as specialized articles, essays and literary works, including prose works from earlier periods recognized as masterpieces.

(Choose one)

Agree

Disagree

38. I can read reasonably legible handwriting without difficulty.

(Choose one)

Agree

Disagree

39. How many years of formal English education have you had?

40. At work, how often do you:

41. read English-language text?

(Choose one)

Daily

Weekly

Monthly

Less Often

I don't read English-language text at work.

42. write English-language text?

(Choose one)

Every day or almost every day

Once or twice a week

Once or twice per month

Less often, but every once in a while

I never write English-language text at work.

43. speak in English?

(Choose one)

- Every day or almost every day
- Once or twice a week
- Once or twice per month
- Less often, but every once in a while
- I never need to speak English at work.

44. understand spoken English?

(Choose one)

- Every day or almost every day
- Once or twice a week
- Once or twice per month
- Less often, but every once in a while
- I don't need to understand spoken English at work.

45. When you use electronic resources to access medical information, is that information

(Choose one)

- Mostly in English
- Mostly in Spanish
- About equal
- I don't use electronic resources to access medical information

46. Have you ever used dictionary or translation web sites (Google Translate, Babelfish, etc.)?

(Choose one)

Yes

No

47. How often do you use such websites?

(Choose one)

Every day or almost every day

Once or twice a week

Once or twice per month

Less often, but every once in a while

48. Is there anything else you wish to say about how you or your colleagues use the English language in your workplace?

Appendix H

Survey Instruments: Language, Spanish

0. Estas preguntas son acerca de sus experiencias con el idioma escrito y hablado en especial el idioma Inglés. Sus respuestas en esta sección son muy importantes y útiles, pero si hay preguntas que no desea contestar, puede saltarlas.

1. ¿Cual de los siguientes es el idioma que habla en su casa?

(Opción múltiple)

Español

Portugués

Inglés

Francés

Italiano

Quechua

Guaraní

Aymara

Otro

2. Otro

3. ¿Cual de los siguientes es el idiomas que utiliza regularmente en el trabajo?

(Opción múltiple)

Español

Portugués

Inglés

Francés

Italiano

Quechua

Guaraní

Aymara

Otro

4. Otro

5. Las siguientes preguntas son sobre como piense de su nivel de competencia con la idioma Inglés. Segun su nivel competencia en el tema con una escala de 1-5, siendo 1 no competencia y 5 experto de el dominio.

6. Respecto al uso del idioma Inglés en contextos no médicos, ¿como calificaría usted su nivel de dominio de:

7. lectura en Inglés?

(Elege uno)

1 - No sabe leer Inglés

2

3

4

5 - Capacidad de lectura en inglés perfecta

8. escribir Inglés?

(Elege uno)

1 - No sabe escribir nada en Inglés

2

3

4

5 - Capacidad de la escritura perfecta

9. hablar Inglés?

(Elege uno)

1 - No sabe hablar Inglés

2

3

4

5 - Capacidad de hablar perfecta

10. escuchar Inglés?

(Elege uno)

1 - No entiende cuando le hablan en Inglés

2

3

4

5 - Entiende perfecto el idioma Inglés cuando le hablan

11. Respecto al uso del ingles en contextos médicos, ¿como calificaría usted su nivel de dominio de:

12. lectura en Inglés?

(Elege uno)

1 - No sabe leer Inglés

2

3

4

5 - Capacidad de lectura en inglés perfecta

13. escribir Inglés?

(Elege uno)

1 - No sabe escribir nada en Inglés

2

3

4

5 - Capacidad de la escritura perfecta

14. hablar Inglés?

(Elege uno)

1 - No sabe hablar Inglés

2

3

4

5 - Capacidad de hablar perfecta

15. escuchar Inglés?

(Elege uno)

1 - No entiende cuando le hablan en Inglés

2

3

4

5 - Entiende perfecto el idioma Inglés cuando le hablan

16. En comparación con sus colegas, piensa usted que su nivel general de competencia en Inglés es

(Elege uno)

Superior a la media

Igual a la media

Inferior a la media

17. Marque "de acuerdo" o "en desacuerdo" al lado de las siguientes afirmaciones acerca de su habilidad para leer textos en idioma Inglés.

18. Puedo reconocer e indentificar todas las letras del alfabeto.

(Elege uno)

De acuerdo

En desacuerdo

19. Puedo leer algunas palabras y frases aisladas, como los números y nombres comunes, que veo en las señales, los menús, y tiendas, y en materiales sencillos y cotidianos como anuncios y publicidades.

(Elege uno)

De acuerdo

En desacuerdo

20. Puedo entender el propósito y el significado principal de textos muy breves y simples, como notas personales escritas, anuncios comerciales, anuncios públicos, mapas, etc.

(Elege uno)

De acuerdo

En desacuerdo

21. Puedo entender instrucciones sencillas, por ejemplo como llegar a un determinado lugar.

(Elege uno)

De acuerdo

En desacuerdo

22. Puedo entender textos, cortos y simples sobre personas, lugares, y cosas conocidos como los que se encuentran en folletos turísticos.

(Elege uno)

De acuerdo

En desacuerdo

23. Puedo entender textos fácticos y sencillos, tales como noticias breves del diario , datos biográficos, descripciones o material técnico simple.

(Elege uno)

De acuerdo

En desacuerdo

24. Puedo entender la idea principal y algunos detalles de textos breves ,sencillos claramente organizados sobre personas, lugares y eventos de los cuales que estoy familiarizado.

(Elege uno)

De acuerdo

En desacuerdo

25. Puedo entender los informes simples y claros sobre eventos actuales y pasados.

(Elege uno)

De acuerdo

En desacuerdo

26. Puedo entender correspondencias sencillas, escritas a máquina en contextos familiares, incluyendo la descripción de eventos, sentimientos, deseos y planes futuros.

(Elege uno)

De acuerdo

En desacuerdo

27. Generalmente puedo entender las ideas principales de la prosa auténtica sobre temas que conozco, ya sea porque pertenecen a mi experiencia laboral o temas que me interesan.

(Elege uno)

De acuerdo

En desacuerdo

28. Por lo general puedo leer y entender todo el material de un periódico de mayor circulación en la ciudad o país con el que estoy familiarizado.

(Elege uno)

De acuerdo

En desacuerdo

29. En la lectura de un periódico o revista que contiene contenido editorial o de opinión, puedo leer "el entre las líneas" y comprender lo que no está directamente declarado.

(Elege uno)

De acuerdo

En desacuerdo

30. Puedo entender la intención del autor y seguir la línea de razonamiento en los textos que incluyen hipótesis, persuasión, opinión justificada (por ejemplo, editoriales, debates y otros artículos de opinión), con poco o ningún uso de un diccionario.

(Elege uno)

De acuerdo

En desacuerdo

31. Puedo entender ensayos contemporáneos y la prosa literaria reciente con poco o ningún uso de un diccionario.

(Elege uno)

De acuerdo

En desacuerdo

32. Puedo entender las ideas principales y los detalles importantes de casi todos los materiales escritos dentro de mi área profesional o área de interés principal (por ejemplo, informes, análisis, cartas, argumentos, etc.).

(Elege uno)

De acuerdo

En desacuerdo

33. capaz de leer con fluidez y precisión todos los estilos y formas del lenguaje pertinentes a las necesidades profesionales o de interés personal sin referencia a un diccionario.

(Elege uno)

De acuerdo

En desacuerdo

34. Comprendo análisis largos y complejos, informes fácticos, y textos literarios.

(Elege uno)

De acuerdo

En desacuerdo

35. Puedo entender el significado y la intención de la mayoría de los usos de los modismos, las referencias culturales, juegos de palabras, el sarcasmo y la ironía, incluso en textos muy abstractos y culturalmente "cargados".

(Elege uno)

De acuerdo

En desacuerdo

36. Puedo entender el lenguaje que ha sido especialmente adaptado para diferentes situaciones, audiencias u objetivos, tales como un ensayo político, una anécdota humorística, broma, sermón, o críticas, y puedo apreciar las diferencias de estilo.

(Elege uno)

De acuerdo

En desacuerdo

37. Soy prácticamente capaz de leer todas las formas de la lengua escrita, incluyendo artículos lingüísticamente complejos como artículos especializados, ensayos y obras literarias, incluyendo las obras en prosa de períodos anteriores reconocidas como obras maestras.

(Elege uno)

De acuerdo

En desacuerdo

38. Puedo leer la letra razonablemente legible sin dificultad.

(Elege uno)

De acuerdo

En desacuerdo

39. ¿Cuántos años de educación formal de Inglés ha tenido?

40. **Respecto al uso del idioma Inglés a su trabajo, ¿con qué frecuencia:**

41. lee textos en idioma Inglés?

(Elege uno)

Daily

Weekly

Monthly

Less Often

Nunca leo texto en idioma Inglés a trabajo.

42. escribe textos en idioma Inglés?

(Elege uno)

Todos los días o casi todos los días

1 o 2 veces por semana

1 o 2 veces por mes

Menos de una vez por mes

No tengo que escribir texto en idioma Inglés.

43. habla el idioma Inglés?

(Elege uno)

- Todos los días o casi todos los días
- 1 o 2 veces por semana
- 1 o 2 veces por mes
- Menos de una vez por mes
- Nunca hablo Inglés.

44. entiende el idioma Inglés hablado?

(Elege uno)

- Todos los días o casi todos los días
- 1 o 2 veces por semana
- 1 o 2 veces por mes
- Menos de una vez por mes
- Nunca entiendo el Inglés hablado.

45. La información médica a la que accedí por los recursos electrónicos es

(Elege uno)

- Mayormente en Inglés
- Mayormente en Español
- En los dos idiomas por igual
- No utilizo los recursos electrónicos para acceder a la información médica

46. ¿Alguna vez has utilizado diccionarios electrónicos o los sitios web de traducción (Google Translate, Babelfish, etc)?

(Elege uno)

Si

No

47. ¿Con qué frecuencia utiliza este tipo de herramientas?

(Elege uno)

Todos los días o casi todos los días

1 o 2 veces por semana

1 o 2 veces por mes

Menos de una vez por mes

48. ¿Hay algo más que nos quiera decir sobre el uso del idioma inglés en su trabajo?

Appendix I

Survey Instruments: Post-task follow-up, English

0. This is the last step in the study. In this step, we would like you to tell us about your experience with the different systems you just used. The three systems were an English-only system, a Spanish-only system, and a bilingual system that included both English and Spanish. Your answers are very important and helpful, but you should feel free to skip any questions that you don't want to answer.

1. I thought that the

(Choose one)

English-only

Spanish-only

Bilingual

system was the easiest one to use.

2. The

(Choose one)

English-only

Spanish-only

Bilingual

system was the most efficient one to use.

3. I was fastest at using the

(Choose one)

English-only

Spanish-only

Bilingual

system.

4. I liked using the

(Choose one)

English-only

Spanish-only

Bilingual

system the most.

5. The

(Choose one)

English-only

Spanish-only

Bilingual

system was the most difficult one to use.

6. What did you find easiest about using the English-only system?

7. What did you find most difficult about using the English-only system?

8. What did you find easiest about using the Spanish-only system?

9. What did you find most difficult about using the Spanish-only system?

10. What did you find easiest about using the Bilingual system?

11. What did you find most difficult about using the Bilingual system?

Appendix J

Survey Instruments: Post-task follow-up, Spanish

0. Este es el último paso en el estudio. En este paso, nos gustaría que nos cuente acerca de su experiencia con los diferentes sistemas que acaba de utilizar: sistema solo en inglés, el sistema solo en español y el sistema bilingüe (español-inglés) Sus respuestas son muy importantes y útiles, pero usted debe sentirse libre de saltar cualquier pregunta que usted no quiere responder.

1. Pienso que el sistema

(Elege uno)

Inglés

Español

Bilingüe

fue el que me resulto más fácil de usar.

2. El sistema

(Elege uno)

Inglés

Español

Bilingüe

fue el que me resulto más eficiente.

3. El sistema

(Elege uno)

Inglés

Español

Bilingüe

fue el más rápido de los tres sistemas usados.

4. De los tres sistemas, el

(Elege uno)

Inglés

Español

Bilingüe

es el que prefiero.

5. El sistema

(Elege uno)

Inglés

Español

Bilingüe

fue el me resulto más difícil de usar.

6. ¿Qué es lo que resultó más fácil al usar el sistema Inglés?

7. ¿Qué es lo que resultó más difícil al usar el sistema Inglés?

8. ¿Qué es lo que resultó más fácil del uso el sistema Español?

9. ¿Qué es lo que resultó más difícil del uso el sistema Español?

10. ¿Qué es lo que resultó más fácil del uso el sistema Bilingüe?

11. ¿Qué es lo que resultó más difícil del uso el sistema Bilingüe?

Appendix K

Subject Recruitment Letter

Note: This is the letter we sent to all prospective subjects. The wording was left unchanged, with the exception of the study participation URL (link), which changed depending on which particular group of subjects for whom the letter was intended.

Estimado colega:

Mi nombre es Steven Bedrick, me desempeño como estudiante de Informática médica en la Oregon Health and Science University (OHSU) donde trabajo con los doctores Nancy Carney y William Hersh.

He tenido el privilegio de conocer algunos de los países latinoamericanos durante mis viajes con la Dra. Carney por proyectos de investigación. Y tras los mismos, decidí enfocar mi tesis doctoral en un tópico de la informática transcultural.

El motivo de esta carta es informarle del estudio e invitarlo a participar en él. Con esta investigación estamos intentando aprender como interfaces multilingües pueden influenciar (mejorar o entorpecer) la habilidad de los usuarios con los motores de búsqueda para recolectar e identificar resultados pertinentes.

Los interesados en participar del estudio deberán completar un cuestionario para luego realizar una serie de tareas simulando una búsqueda de información.

La realización de esta tarea le llevará aproximadamente media hora, pero para algunos participantes este tiempo puede ser menor. Al finalizar el estudio tendrá la posibilidad de descargar un certificado de agradecimiento.

Cualquier personal de la salud puede participar, por ejemplo estudiantes, enfermeras, médicos residentes o especialistas.

El estudio fue aprobado por el comité de ética de mi institución (Institutional Review Board of Oregon Health and Science University (protocol number 5872)). La participación en este estudio será completamente confidencial y no se solicitará en ningún momento ningún tipo de información que pueda identificar a los participantes.

Si usted está interesado en participar solo debe ir a la siguiente dirección de internet:

<http://skynet.ohsu.edu/busctrad/c/nc>

Por favor reenvíe este correo invitando a otros colegas que puedan estar interesados, ya que cuantos más profesionales realicen la encuesta, mejores serán los resultados finales. De esta misma forma Ud. puede colaborar si incluso ha decidido no participar.

Si usted tiene alguna pregunta por favor envíenos un mail a cualquiera de nuestras direcciones de correo electrónico. Gracias!

– Steven Bedrick

Doctoral Candidate, OHSU (bedricks@ohsu.edu)

— Dr. William Hersh, MD

Professor and Chair, Biomedical Informatics and Clinical Epidemiology (hersh@ohsu.edu)

Bibliography

- [1] Curley SP, Connelly DP, Rich EC. Physicians' use of medical knowledge resources: preliminary theoretical framework and findings. *Med Decis Making*. 1990 Jan;10(4):231–41. Available from: http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=2122168&dopt=abstractplus.
- [2] Verhoeven AA, Boerma EJ, de Jong BM. Use of information sources by family physicians: a literature survey. *Bull Med Libr Assoc*. 1995 Jan;83(1):85–90. Available from: http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=7703946&dopt=abstractplus.
- [3] Guyatt G, Rennie D, Meade M, Cook D, editors. *Users' guides to the medical literature: a manual for evidence-based clinical practice*. 2nd ed. New York: McGraw Hill Medical; 2008. Available from: <http://www.loc.gov/catdir/toc/ecip085/2007047778.html>.
- [4] Ely JW. Why can't we answer our questions? *J Fam Pract*. 2001 Nov;50(11):974–5. Available from: <http://www.jfponline.com/Pages.asp?AID=2373>.
- [5] Richardson WS, Wilson MC, Nishikawa J, Hayward RS. The well-built clinical question: a key to evidence-based decisions. *ACP J Club*. 1995 Jan;123(3):A12–3. Available from: http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=7703946&dopt=abstractplus.

nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=7582737&dopt=abstractplus.

- [6] Hersh WR, Hickam DH. How well do physicians use electronic information retrieval systems? A framework for investigation and systematic review. *JAMA*. 1998 Oct;280(15):1347–52. Available from: <http://jama.ama-assn.org/cgi/pmidlookup?view=long&pmid=9794316>.
- [7] Hoogendam A, Stalenhoef AFH, de Vries Robbé PF, Overbeke AJPM. Analysis of queries sent to PubMed at the point of care: observation of search behaviour in a medical teaching hospital. *BMC Med Inform Decis Mak*. 2008 Jan;8:42. Available from: <http://www.biomedcentral.com/1472-6947/8/42>.
- [8] Loria A, Arroyo P. Language and country preponderance trends in MEDLINE and its causes. *Journal of the Medical Library Association : JMLA*. 2005 Jul;93(3):381–5. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/16059428?dopt=abstract>.
- [9] Wolfram|Alpha. Spanish - Wolfram|Alpha;. Available from: <http://www.wolframalpha.com/input/?i=Spanish>.
- [10] Bureau UC. United States - Language Spoken At Home;. Available from: http://factfinder.census.gov/servlet/STTable?_bm=y&-geo_id=01000US&-qr_name=ACS_2009_1YR_G00_S1601&-ds_name=ACS_2009_1YR_G00_-&-_lang=en&-redoLog=false.
- [11] Zeng-Treitler Q, Kim H, Rosemblat G, Keselman A. Can multilingual machine translation help make medical record content more comprehensible to patients? *STUDIES IN HEALTH TECHNOLOGY AND INFORMAT-*

- ICS. 2010 Jan;160(Pt 1):73–7. Available from: <http://booksonline.iospress.nl/Content/View.aspx?piid=17329>.
- [12] Fontelo P, Liu F, Leon S, Anne A, Ackerman M. PICO Linguist and BabelMeSH: development and partial evaluation of evidence-based multi-language search tools for MEDLINE/PubMed. *Medinfo*. 2007 Jan;12(Pt 1):817–21. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/17911830?dopt=abstract>.
- [13] Hersh WR, Donohoe LC. SAPHIRE International: a tool for cross-language information retrieval. *Proceedings / AMIA Annual Symposium AMIA Symposium*. 1998 Jan;p. 673–7. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/9929304?dopt=abstract>.
- [14] Biblioteca Virtual em Saúde; [cited January 1, 2010]. Available from: <http://regional.bvsalud.org/php/index.php>.
- [15] Ammon U, editor. *The dominance of English as a language of science: effects on other languages and language communities*. vol. 84. Berlin: Mouton de Gruyter; 2001. Available from: <http://www.loc.gov/catdir/toc/fy032/2001030388.html>.
- [16] Kaplan RB. English: the Accidental Language of Science? In: Ammon U, editor. *The dominance of English as a language of science: effects on other languages and language communities*. vol. 84. Berlin: Mouton de Gruyter; 2001. Available from: <http://www.loc.gov/catdir/toc/fy032/2001030388.html>.
- [17] Wong ML. On science and English. *EMBO Rep*. 2007 Apr;8(4):302. Available from: <http://www.nature.com/doifinder/10.1038/sj.embor.7400946>.

- [18] Monge-Nájera J, Nielsen V. The countries and languages that dominate biological research at the beginning of the 21st century. *Rev Biol Trop*. 2005 Jan;53(1-2):283–94. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/17354441?dopt=abstract>.
- [19] Moher D, Fortin P, Jadad AR, Jüni P, Klassen T, Lorier JL, et al. Completeness of reporting of trials published in languages other than English: implications for conduct and reporting of systematic reviews. *Lancet*. 1996 Feb;347(8998):363–6. Available from: http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=8598702&dopt=abstractplus.
- [20] Moher D, Pham B, Lawson ML, Klassen TP. The inclusion of reports of randomised trials published in languages other than English in systematic reviews. *Health technology assessment (Winchester, England)*. 2003 Jan;7(41):1–90. Available from: <http://www.hta.ac.uk/execsumm/summ741.htm>.
- [21] Clark OAC, Castro AA. Searching the Literatura Latino Americana e do Caribe em Ciências da Saúde (LILACS) database improves systematic reviews. *International journal of epidemiology*. 2002 Feb;31(1):112–4. Available from: <http://ije.oxfordjournals.org/cgi/content/full/31/1/112>.
- [22] Fung IC. Seek, and ye shall find: Accessing the global epidemiological literature in different languages. *Emerging themes in epidemiology*. 2008 Jan;5:21. Available from: <http://www.ete-online.com/content/5/1/21>.
- [23] Muccioli C, Campos M, Goldchmit M, Dantas PEC, Bechara SJ, Costa VP. [Articles in English in the Brazilian Archives of Ophthalmology:

- a result of globalization]. *Arquivos brasileiros de oftalmologia*. 2006 Jan;69(4):461. Available from: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0004-27492006000400001&lng=en&nrm=iso&tlng=en.
- [24] Aleixandre-Benavent R, Zurián JCV, Alonso-Arroyo A, Miguel-Dasit A, de Dios JG, de Granda Orive J. Español frente a inglés como idioma de publicación y factor de impacto de NEUROLOGÍA. *Neurología* (Barcelona, Spain). 2007 Jan;22(1):19–26. Available from: <http://www.arsxxi.com/Revistas/mostrarticulo.php?idarticulo=7109114>.
- [25] Egger M, Zellweger-Zähner T, Schneider M, Junker C, Lengeler C, Antes G. Language bias in randomised controlled trials published in English and German. *Lancet*. 1997 Aug;350(9074):326–9. Available from: [http://linkinghub.elsevier.com/retrieve/pii/S0140-6736\(97\)02419-7](http://linkinghub.elsevier.com/retrieve/pii/S0140-6736(97)02419-7).
- [26] Meneghini R, Packer AL. Is there science beyond English? Initiatives to increase the quality and visibility of non-English publications might help to break down language barriers in scientific communication. *EMBO Rep*. 2007 Feb;8(2):112–6. Available from: <http://www.nature.com/doifinder/10.1038/sj.embor.7400906>.
- [27] Man JP, Weinkauff JG, Tsang M, Sin DD. Why do some countries publish more than others? An international comparison of research funding, English proficiency and publication output in highly ranked general medical journals. *Eur J Epidemiol*. 2004 Jan;19(8):811–7. Available from: <http://www.springerlink.com/content/q426464q754g3386/>.

- [28] Vasconcelos SMR, Sorenson MM, Leta J. Scientist-friendly policies for non-native English-speaking authors: timely and welcome. *Braz J Med Biol Res.* 2007 Jun;40(6):743–7. Available from: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-879X2007000600001&lng=en&nrm=iso&tlng=en.
- [29] Link AM. US and non-US submissions: an analysis of reviewer bias. *JAMA.* 1998 Jul;280(3):246–7. Available from: <http://jama.ama-assn.org/cgi/content/full/280/3/246>.
- [30] Coates R, Sturgeon B, Bohannan J, Pasini E. Language and publication in "Cardiovascular Research" articles. *Cardiovasc Res.* 2002 Feb;53(2):279–85. Available from: <http://cardiovascres.oxfordjournals.org/cgi/content/full/53/2/279>.
- [31] Victora CG, Moreira CB. [North-South relations in scientific publications: editorial racism?]. *Revista de saúde pública.* 2006 Aug;40 Spec no.:36–42. Available from: <http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=ShowDetailView&TermToSearch=16924301>.
- [32] Gannon F. Language barriers. *EMBO Rep.* 2008 Mar;9(3):207. Available from: <http://www.nature.com/doifinder/10.1038/embor.2008.14>.
- [33] Horton R. North and South: bridging the information gap. *Lancet.* 2000 Jun;355(9222):2231–6. Available from: [http://linkinghub.elsevier.com/retrieve/pii/S0140-6736\(00\)02414-4](http://linkinghub.elsevier.com/retrieve/pii/S0140-6736(00)02414-4).
- [34] Fernández E, García AM. Accuracy of referencing of Spanish names in Medline. *Lancet.* 2003 Jan;361(9354):351–2. Available from: [http://linkinghub.elsevier.com/retrieve/pii/S0140-6736\(03\)12356-2](http://linkinghub.elsevier.com/retrieve/pii/S0140-6736(03)12356-2).

- [35] Ruiz-Pérez R, López-Cózar ED, Jiménez-Contreras E. Spanish name indexing errors in international databases. *Lancet*. 2003 May;361(9369):1656–7. Available from: [http://linkinghub.elsevier.com/retrieve/pii/S0140-6736\(03\)13285-0](http://linkinghub.elsevier.com/retrieve/pii/S0140-6736(03)13285-0).
- [36] Navarro FA, Barnes J. Traducción de títulos al inglés en MEDICINA CLÍNICA: calidad e influencia del castellano [Translation of titles into English in Medicina Clínica: quality and influence of the Spanish language]. *Medicina clínica*. 1996 Mar;106(8):298–303. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/8667688?dopt=abstract>.
- [37] Meneghini R, Packer AL, Nassi-Calò L. Articles by latin american authors in prestigious journals have fewer citations. *PLoS ONE*. 2008 Jan;3(11):e3804. Available from: <http://www.plosone.org/article/info%253Adoi%252F10.1371%252Fjournal.pone.0003804>.
- [38] Albert T. Scientific communication—not only in English. *Lancet*. 2001 Oct;358(9291):1388. Available from: [http://linkinghub.elsevier.com/retrieve/pii/S0140-6736\(01\)06493-5](http://linkinghub.elsevier.com/retrieve/pii/S0140-6736(01)06493-5).
- [39] Vasconcelos SMR, Sorenson MM, Leta J, Sant’ana M, Batista PD. Researchers’ writing competence: a bottleneck in the publication of Latin-American science? *EMBO Rep*. 2008 Aug;9(8):700–2. Available from: <http://www.nature.com/doifinder/10.1038/embor.2008.143>.
- [40] Charlton BG. How can the English-language scientific literature be made more accessible to non-native speakers? Journals should allow greater use of referenced direct quotations in ‘component-oriented’ scientific writing. *Med Hypotheses*. 2007 Jan;69(6):1163–4. Available from: [http://linkinghub.elsevier.com/retrieve/pii/S0306-9877\(07\)00449-5](http://linkinghub.elsevier.com/retrieve/pii/S0306-9877(07)00449-5).

- [41] Benfield JR, Feak CB. How authors can cope with the burden of English as an international language. *Chest*. 2006 Jun;129(6):1728–30. Available from: <http://www.chestjournal.org/cgi/content/full/129/6/1728>.
- [42] Freeman P, Robbins A. The publishing gap between rich and poor: the focus of AuthorAID. *Journal of public health policy*. 2006 Jul;27(2):196–203. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/16961198?dopt=abstract>.
- [43] Lanza V. Some notes on medical English. *Journal of clinical monitoring and computing*. 2005 Jun;19(3):179–81. Available from: <http://www.springerlink.com/content/b46228575085q4m5/>.
- [44] Kirkman J. Writing in English for an international readership. *BMJ*. 1996 Nov;313(7068):1321–2; discussion 1323. Available from: <http://www.bmj.com/cgi/content/full/313/7068/1321/a>.
- [45] Tompson A. How to write an English medical manuscript that will be published and have impact. *Surg Today*. 2006 Jan;36(5):407–9. Available from: <http://www.springerlink.com/content/k38g821313114503/>.
- [46] Fung IC. Open access for the non-English-speaking world: overcoming the language barrier. *Emerging themes in epidemiology*. 2008 Jan;5:1. Available from: <http://www.ete-online.com/content/5/1/1>.
- [47] Bordons M. Hacia el reconocimiento internacional de las publicaciones científicas españolas. *Rev Esp Cardiol*. 2004;57(9):799–802.
- [48] Ospina EG, Herault LR, Cardona AF. Uso de bases de datos bibliográficas por investigadores biomédicos latinoamericanos hispanoparlantes: estudio transversal [The use of bibliographic databases by

- Spanish-speaking Latin American biomedical researchers: a cross-sectional study]. *Rev Panam Salud Publica*. 2005 Apr;17(4):230–6. Available from: <http://www.ingentaconnect.com/content/paho/pajph/2005/00000017/00000004/art00003?token=00571d08227e71de114486e5865462440444255474621583f3b25535e4e26634a4>
- [49] World Health Organization. WHO HINARI Access to Research in Health Programme;. Available from: <http://www.who.int/hinari/en/>.
- [50] Packer A, Biojone M, Antonio I, Takenaka R, Garcia A, da Silva A, et al. SciELO: uma metodologia para publicação eletrônica. *Ciência Informação*. 1998;27:109–121. Available from: <http://revista.ibict.br/index.php/ciinf/article/viewArticle/339>.
- [51] Blank D, Buchweitz C, Procianoy RS. Impact of SciELO and MEDLINE indexing on submissions to *Jornal de Pediatria*. *J Pediatr (Rio J)*. 2005 Jan;81(6):431–4. Available from: http://www.jpmed.com.br/conteudo/Ing_resumo.asp?varArtigo=1414&cod=&idSecao=4.
- [52] Curioso WH, Arriola-Quiroz I, Cruz-Encarnación M. [A simple strategy to improve searching of indexed articles in SciELO]. *Revista médica de Chile*. 2008 May;136(6):812–4. Available from: <http://www.scielo.cl/cgi-bin/fbpe/fbtext?pid=S0034-98872008000600020&lng=en&nrm=iso&tlng=en>.
- [53] Hodgkin K. Diagnostic vocabulary for primary care. *J Fam Pract*. 1979 Jan;8(1):129–44. Available from: http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=759538&dopt=abstractplus.

- [54] Pauker SG, Gorry GA, Kassirer JP, Schwartz WB. Towards the simulation of clinical cognition. Taking a present illness by computer. *Am J Med.* 1976 Jun;60(7):981–96.
- [55] González-González AI, Dawes M, Sánchez-Mateos J, Riesgo-Fuertes R, Escortell-Mayor E, Sanz-Cuesta T, et al. Information needs and information-seeking behavior of primary care physicians. *Annals of family medicine.* 2007 Jan;5(4):345–52. Available from: <http://www.annfammed.org/cgi/content/full/5/4/345>.
- [56] Smith R. What clinical information do doctors need? *BMJ.* 1996 Oct;313(7064):1062–8. Available from: <http://www.bmj.com/cgi/content/full/313/7064/1062?view=long&pmid=8898602>.
- [57] Wyatt J. Use and sources of medical knowledge. *Lancet.* 1991 Nov;338(8779):1368–73. Available from: http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=1682745&dopt=abstractplus.
- [58] Ziman JM. The proliferation of scientific literature: a natural process. *Science.* 1980 Apr;208(4442):369–71. Available from: <http://www.sciencemag.org/cgi/reprint/208/4442/369>.
- [59] Case DO. Information Behavior: An Introduction. In: Looking for information: a survey of research on information seeking, needs, and behavior. 2nd ed. Library and information science. Amsterdam: Elsevier/Academic Press; 2007. Available from: <http://www.loc.gov/catdir/enhancements/fy0702/2006050838-d.html>.
- [60] Case DO. Information Needs and Information Seeking. In: Looking for information: a survey of research on information seeking, needs, and be-

- havior. 2nd ed. Library and information science. Amsterdam: Elsevier/Academic Press; 2007. Available from: <http://www.loc.gov/catdir/enhancements/fy0702/2006050838-d.html>.
- [61] Belkin NJ. Anomalous State of Knowledge. In: Fisher KE, Erdelez S, McKechnie L, editors. Theories of information behavior. Medford, N.J.: Published for the American Society for Information Science and Technology by Information Today; 2005. p. 44–48. Available from: <http://www.loc.gov/catdir/toc/ecip0511/2005010420.html>.
- [62] Dervin B, Foreman-Wernet L, Lauterbach E. Sense-making methodology reader: selected writings of Brenda Dervin. Cresskill, N.J.: Hampton Press; 2003.
- [63] Dervin B. Information as a user construct: The relevance of perceived information needs to synthesis and interpretation. In: Ward SA, Reed LJ, editors. Knowledge structure and use: Implications for synthesis and interpretation. Temple University Press; 1983. p. 153–184.
- [64] Ingwersen P, Jarvelin K. The turn: integration of information seeking and retrieval in context. Kluwer international series on information retrieval. Dordrecht: Springer; 2005. Available from: <http://www.loc.gov/catdir/enhancements/fy0901/2008422356-d.html>.
- [65] Wilson TD. Evolution in Information Behavior Modeling: Wilson's Model. In: Fisher KE, Erdelez S, McKechnie L, editors. Theories of information behavior. Medford, N.J.: Published for the American Society for Information Science and Technology by Information Today; 2005. Available from: <http://www.loc.gov/catdir/toc/ecip0511/2005010420.html>.

- [66] Case DO. Models of Information Behavior. In: Looking for information: a survey of research on information seeking, needs, and behavior. 2nd ed. Library and information science. Amsterdam: Elsevier/Academic Press; 2007. Available from: <http://www.loc.gov/catdir/enhancements/fy0702/2006050838-d.html>.
- [67] Leckie GJ, Pettigrew KE, Sylvain C. Modeling the information seeking of professionals: A general model derived from research on engineers, health care professionals and lawyers. *Library Quarterly*. 1996;66:161–193.
- [68] Marchionini G. Information seeking in electronic environments. Cambridge: Cambridge University Press; 1995.
- [69] Case DO. Looking for information: a survey of research on information seeking, needs, and behavior. 2nd ed. Library and information science. Amsterdam: Elsevier/Academic Press; 2007. Available from: <http://www.loc.gov/catdir/enhancements/fy0702/2006050838-d.html>.
- [70] Fisher KE, Erdelez S, McKechnie L, editors. Theories of information behavior. Medford, N.J.: Published for the American Society for Information Science and Technology by Information Today; 2005. Available from: <http://www.loc.gov/catdir/toc/ecip0511/2005010420.html>.
- [71] Hearst M. Search user interfaces. Cambridge: Cambridge University Press; 2009. Available from: <http://searchuserinterfaces.com/book>.
- [72] Thompson ML. Characteristics of information resources preferred by primary care physicians. *Bull Med Libr Assoc*. 1997 Apr;85(2):187–92. Available from: http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=9160156&dopt=abstractplus.

- [73] Henefer J, Fulton C. Krikelas's Model of Information Seeking. In: Fisher KE, Erdelez S, McKechnie L, editors. *Theories of information behavior*. Medford, N.J.: Published for the American Society for Information Science and Technology by Information Today; 2005. Available from: <http://www.loc.gov/catdir/toc/ecip0511/2005010420.html>.
- [74] Bryant SL. The information needs and information seeking behaviour of family doctors. *Health Info Libr J*. 2004 Jun;21(2):84–93. Available from: <http://onlinelibrary.wiley.com/doi/10.1111/j.1471-1842.2004.00490.x/abstract>.
- [75] Chambliss ML, Conley J. Answering clinical questions. *J Fam Pract*. 1996 Aug;43(2):140–4. Available from: http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=8708623&dopt=abstractplus.
- [76] Gorman PN, Ash J, Wykoff L. Can primary care physicians' questions be answered using the medical journal literature? *Bulletin of the Medical Library Association*. 1994 Apr;82(2):140–6. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/7772099?dopt=abstract>.
- [77] Gorman PN, Helfand M. Information seeking in primary care: how physicians choose which clinical questions to pursue and which to leave unanswered. *Medical decision making : an international journal of the Society for Medical Decision Making*. 1995 Jan;15(2):113–9. Available from: <http://mdm.sagepub.com/cgi/reprint/15/2/113>.
- [78] Gorman P. Information needs of physicians. *Journal of the American Society for Information Science*. 1995;46(10):729–736.

- [79] Ely JW, Osheroff JA, Ebell MH, Bergus GR, Levy BT, Chambliss ML, et al. Analysis of questions asked by family doctors regarding patient care. *BMJ*. 1999 Aug;319(7206):358–61. Available from: http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=10435959&dopt=abstractplus.
- [80] Ely JW, Osheroff JA, Gorman PN, Ebell MH, Chambliss ML, Pifer EA, et al. A taxonomy of generic clinical questions: classification study. *BMJ*. 2000 Aug;321(7258):429–32. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/10938054?dopt=abstract>.
- [81] Coumou HCH, Meijman FJ. How do primary care physicians seek answers to clinical questions? A literature review. *Journal of the Medical Library Association : JMLA*. 2006 Jan;94(1):55–60. Available from: http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=16404470&dopt=abstractplus.
- [82] Osheroff JA, Forsythe DE, Buchanan BG, Bankowitz RA, Blumenfeld BH, Miller RA. Physicians' information needs: analysis of questions posed during clinical teaching. *Ann Intern Med*. 1991 Apr;114(7):576–81. Available from: http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=2001091&dopt=abstractplus.
- [83] Rosenberg W, Donald A. Evidence based medicine: an approach to clinical problem-solving. *BMJ*. 1995 Apr;310(6987):1122–6. Available from: <http://www.bmj.com/content/310/6987/1122.long>.
- [84] Horsley T, O'Neill J, McGowan J, Perrier L, Kane G, Campbell C. Interventions to improve question formulation in professional practice and self-directed learning. *Cochrane Database Syst Rev*. 2010 Jan;5:CD007335.

Available from: <http://onlinelibrary.wiley.com/o/cochrane/clsysrev/articles/CD007335/frame.html>.

- [85] Bennett NL, Casebeer LL, Kristofco R, Collins BC. Family physicians' information seeking behaviors: a survey comparison with other specialties. *BMC Med Inform Decis Mak*. 2005 Jan;5:9. Available from: <http://www.biomedcentral.com/1472-6947/5/9>.
- [86] Ely JW, Burch RJ, Vinson DC. The information needs of family physicians: case-specific clinical questions. *J Fam Pract*. 1992 Sep;35(3):265-9. Available from: http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=1517722&dopt=abstractplus.
- [87] Connelly DP, Rich EC, Curley SP, Kelly JT. Knowledge resource preferences of family physicians. *J Fam Pract*. 1990 Mar;30(3):353-9. Available from: http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=2248632&dopt=abstractplus.
- [88] Huth EJ. Needed: an economics approach to systems for medical information. *Ann Intern Med*. 1985 Oct;103(4):617-9. Available from: http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=3929662&dopt=abstractplus.
- [89] Ely JW, Osheroff JA, Chambliss ML, Ebell MH, Rosenbaum ME. Answering physicians' clinical questions: obstacles and potential solutions. *Journal of the American Medical Informatics Association : JAMIA*. 2005 Jan;12(2):217-24. Available from: <http://jamia.bmj.com/content/12/2/217.long>.
- [90] Haug JD. Physicians' preferences for information sources: a meta-analytic study. *Bull Med Libr Assoc*. 1997 Jul;85(3):223-32. Available from:

http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=9285121&dopt=abstractplus.

- [91] Hider PN, Griffin G, Walker M, Coughlan E. The information-seeking behavior of clinical staff in a large health care organization. *Journal of the Medical Library Association : JMLA*. 2009 Jan;97(1):47–50. Available from: http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=19159006&dopt=abstractplus.
- [92] Alper BS, Stevermer JJ, White DS, Ewigman BG. Answering family physicians' clinical questions using electronic medical databases. *J Fam Pract*. 2001 Nov;50(11):960–5. Available from: <http://www.jfponline.com/Pages.asp?AID=2383>.
- [93] Hersh WR, Pentecost J, Hickam D. A task-oriented approach to information retrieval evaluation. *Journal of the American Society for Information Science*. 1996;47(1):50–56.
- [94] Herskovic JR, Tanaka LY, Hersh WR, Bernstam EV. A day in the life of PubMed: analysis of a typical day's query log. *Journal of the American Medical Informatics Association : JAMIA*. 2007 Jan;14(2):212–20. Available from: <http://jamia.bmj.com/content/14/2/212.long>.
- [95] Meats E, Brassey J, Heneghan C, Glasziou P. Using the Turning Research Into Practice (TRIP) database: how do clinicians really search? *Journal of the Medical Library Association : JMLA*. 2007 Apr;95(2):156–63. Available from: http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=17443248&dopt=abstractplus.
- [96] Haynes RB, McKibbon KA, Walker CJ, Ryan N, Fitzgerald D, Ramsden MF. Online access to MEDLINE in clinical settings. A study

- of use and usefulness. *Ann Intern Med.* 1990 Jan;112(1):78–84. Available from: http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=2403476&dopt=abstractplus.
- [97] Westbrook JI, Coiera EW, Gosling AS. Do online information retrieval systems help experienced clinicians answer clinical questions? *Journal of the American Medical Informatics Association : JAMIA.* 2005 Jan;12(3):315–21. Available from: <http://www.jamia.org/cgi/pmidlookup?view=long&pmid=15684126>.
- [98] Ash JS, Berg M, Coiera E. Some unintended consequences of information technology in health care: the nature of patient care information system-related errors. *Journal of the American Medical Informatics Association : JAMIA.* 2004 Jan;11(2):104–12. Available from: <http://jamia.bmj.com/content/11/2/104.long>.
- [99] Davidoff F, Florance V. The informationist: a new health profession? *Ann Intern Med.* 2000 Jun;132(12):996–8. Available from: <http://www.annals.org/content/132/12/996.long>.
- [100] Grefsheim SF, Whitmore SC, Rapp BA, Rankin JA, Robison RR, Canto CC. The informationist: building evidence for an emerging health profession. *Journal of the Medical Library Association : JMLA.* 2010 Apr;98(2):147–56. Available from: http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=20428280&dopt=abstractplus.
- [101] Giuse NB, Koonce TY, Jerome RN, Cahall M, Sathe NA, Williams A. Evolution of a mature clinical informationist model. *Journal of the American*

- Medical Informatics Association : JAMIA. 2005 Jan;12(3):249–55. Available from: <http://jamia.bmj.com/content/12/3/249.long>.
- [102] Giuse NB, Kafantaris SR, Miller MD, Wilder KS, Martin SL, Sathe NA, et al. Clinical medical librarianship: the Vanderbilt experience. *Bulletin of the Medical Library Association*. 1998 Jul;86(3):412–6. Available from: http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=9681179&dopt=abstractplus.
- [103] Martinez-Silveira MS, Oddone N. Information-seeking behavior of medical residents in clinical practice in Bahia, Brazil. *Journal of the Medical Library Association : JMLA*. 2008 Oct;96(4):381–4. Available from: http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=18974818&dopt=abstractplus.
- [104] Nirenburg S. Introduction. In: Nirenburg S, Somers HL, Wilks Y, editors. *Readings in machine translation*. Cambridge, Mass.: MIT Press; 2003. .
- [105] Nirenburg S, Somers HL, Wilks Y, editors. *Readings in machine translation*. Cambridge, Mass.: MIT Press; 2003.
- [106] Automatic Language Processing Advisory Committee. *Languages and Machines: Computers in translation and linguistics*. Division of Behavioral Sciences, National Academy of Sciences, National Research Council; Publication 1416.
- [107] Hutchins J. ALPAC: The (In)Famous Report. *MT News International*. 1996 June;(14):9–12.
- [108] Hutchins W. Machine translation and machine aided translation. *Journal of Documentation*. 1978;34(2):119–159.

- [109] Tucker VM A, Leon M. PAHO Machine Translation System: Introduction and User's Manual. Washington, D.C.; 1980.
- [110] Brown P, Cocke J, Pietra S, Pietra V, Jelinek F, Lafferty J, et al. A statistical approach to machine translation. *Computational Linguistics*. 1990 Jun;16(2). Available from: <http://portal.acm.org/citation.cfm?id=92858.92860>.
- [111] Goutte C, Cancedda N, Dymetman M, Foster G, editors. Learning machine translation. Neural information processing series. Cambridge, Mass.: MIT Press; 2009. Available from: <http://www.loc.gov/catdir/toc/fy0903/2008029324.html>.
- [112] Levow GA, Oard DW, Resnik P. Dictionary-based techniques for cross-language information retrieval. *Information Processing & Management*. 2005;41(3):523 – 547. Cross-Language Information Retrieval. Available from: <http://www.sciencedirect.com/science/article/B6VC8-4D4D0GV-1/2/d61ad5ca339ba923e51adfb2470b3128>.
- [113] Oard DW. Evaluating Interactive Cross-Language Information Retrieval: Document Selection. vol. 2069; 2001. p. 57–71. Available from: <http://www.springerlink.com/content/mnmh96hqd5r4fc1v/>.
- [114] Oard DW. Alternative approaches for cross-language text retrieval; 1997. Available from: mack.ittc.ku.edu/oard97alternative.html.
- [115] Oard D, Resnik P. Support for interactive document selection in cross-language information retrieval. *Information Processing and Management*. 1999;35(3):363–379.

- [116] López-Ostenero F, Gonzalo J, Verdejo F. Noun phrases as building blocks for cross-language search assistance. *Information Processing and Management*. 2005;41(3):549–568.
- [117] Suzuki M, Inoue N, Hashimoto K. A Method for Supporting Document Selection in Cross-language Information Retrieval and its Evaluation. *Computers and the Humanities*. 2001;35:421–438. 10.1023/A:1011877503081. Available from: <http://dx.doi.org/10.1023/A:1011877503081>.
- [118] Petrelli D, Levin S, Beaulieu M, Sanderson M. Which User Interaction for Cross-Language IR? *Design Issues and Reflections*. *J Am Soc Inf Sci*. 2006 Mar;57(5):709–722.
- [119] Capstick J, Erbach G, Uskoreit H. Design and Evaluation of a Psychological Experiment on the Effectiveness of Document Summarisation for the Retrieval of Multilingual WWW Documents. *Working Notes of the AAAI Spring Symposium Intelligent Text Summarisation*. 1998 Feb;.
- [120] Grossman DA, Frieder O. Cross-Language Information Retrieval. In: *Information retrieval: algorithms and heuristics*. vol. SECS 461. Boston: Kluwer; 1998. Available from: <http://www.loc.gov/catdir/enhancements/fy0820/98030435-d.html>.
- [121] Ballesteros L, Croft W. Resolving ambiguity for cross-language retrieval. *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. 1998 Aug; Available from: <http://portal.acm.org/citation.cfm?id=290941.290958>.

- [122] W OD, R DA. Cross-language information retrieval. In: Annual Review of Information Science and Technology. vol. 33. American Society for Information Science.; 1998. .
- [123] Cleverdon C. The Cranfield tests on index language devices. *Aslib Proceedings*. 1963;15(4):106–130.
- [124] Robertson SE, Hancock-Beaulieu MM. On the evaluation of IR systems. *Information Processing & Management*. 1992 Mar;28:457–466. Available from: <http://portal.acm.org/citation.cfm?id=149509.149513>.
- [125] Voorhees E, Harman DK. *TREC: experiment and evaluation in information retrieval*. Cambridge, Mass.: MIT Press; 2005.
- [126] Oard DW, Gonzalo J, Sanderson M, López-Ostenero F, Wang J. Interactive Cross-Language Document Selection. *Information Retrieval*. 2004;7(1-2):205–228.
- [127] Swanson D. Historical note: Information retrieval and the future of an illusion. *Journal of the American Society for Information Science*. 1988;39(2):92–98.
- [128] Hersh WR, Turpin A, Price S, Kraemer D, Olson D, Chan B, et al. Challenging conventional assumptions of automated information retrieval with real users: Boolean searching and batch retrieval evaluations. *Information Processing and Management*. 2001;37(3):383–402.
- [129] Allan J, Carterette B, Lewis J. When will information retrieval be "good enough?"; 2005. p. 433–440. Available from: <http://doi.acm.org/10.1145/1076034.1076109>.

- [130] Smith C, Kantor P. User adaptation: good results from poor systems. SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. 2008 Jul; Available from: <http://portal.acm.org/citation.cfm?id=1390334.1390362>.
- [131] Hersh WR, Turpin A, Price S, Chan B, Kramer D, Sacherek L, et al. Do batch and user evaluations give the same results? Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. 2000;p. 17–24.
- [132] Stead WW, Haynes RB, Fuller S, Friedman CP, Travis LE, Beck JR, et al. Designing medical informatics research and library–resource projects to increase what is learned. *Journal of the American Medical Informatics Association : JAMIA*. 1994 Jan;1(1):28–33. Available from: <http://jamia.bmj.com/content/1/1/28>.
- [133] Marchionini G. Information seeking in full-text end-user-oriented search systems: The roles of domain and search expertise. *Library and Information Science Research*. 1993;15(1):35–69.
- [134] Ruthven I. Interactive information retrieval. *Annual Review of Information Science & Technology*. 2007;42:43–91.
- [135] Ingwersen P. Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory. *Journal of Documentation*. 1996;52(1):3–50.
- [136] Hersh WR. Relevance and retrieval evaluation: perspectives from medicine. *Journal of the American Society for Information Science*. 1994;45(3):201–206.
- [137] Kelly D. Methods for Evaluating Interactive Information Retrieval Systems with Users. *Found Trends Inf Retr*. 2009;3(1–2):1–224.

- [138] Salton G. Evaluation problems in interactive information retrieval. *Information Storage and Retrieval*. 1970;6(1):29 – 44. Available from: <http://www.sciencedirect.com/science/article/B6X2J-465CX28-6R/2/34da755b76d47e3c6d89788094e61298>.
- [139] Borlund P, Ingwersen P. The Development of a Method for the Evaluation of Interactive Information Retrieval Systems. *Journal of Documentation*. 1997;53(3):225—250.
- [140] Borlund P. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research*. 2003;8(3). Available from: <http://informationr.net/ir/8-3/paper152.html>.
- [141] Borlund P. Experimental components for the evaluation of interactive information retrieval systems. *Journal of Documentation*. 2000;56(1):71–90.
- [142] Wang X, Liebscher P, Marchionini G. Improving information seeking performance in hypertext: Roles of display format and search strategy. Center for Automation Research, University of Maryland; 1988. CAR-TR-353.
- [143] Marchionini G, Liebscher P. Performance in electronic encyclopedias: Implications for adaptive systems. In: *Proceedings of the fifty-fourth annual meeting of the ASIS*; 1991. p. 39–48.
- [144] Donabedian A. *Explorations in quality assessment and monitoring*. Ann Arbor, Mich.: Health Administration Press; 1982.
- [145] Hersh WR, Elliot DL, Hickam DH, Wolf SL, Molnar A, Leichtenstien C. Towards new measures of information retrieval evaluation. *Proceedings / the Annual Symposium on Computer Application [sic] in Medical Care Symposium on Computer Applications in Medical Care*. 1994 Jan;p. 895–9.

Available from: http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=7950053&dopt=abstractplus.

- [146] Al-Maskari A. Beyond classical measures: how to evaluate the effectiveness of interactive information retrieval system? SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. 2007 Jul; Available from: <http://portal.acm.org/citation.cfm?id=1277741.1277984>.
- [147] Belkin N. Some(what) grand challenges for information retrieval. SIGIR Forum. 2008 Jun;42(1). Available from: <http://portal.acm.org/citation.cfm?id=1394251.1394261>.
- [148] Järvelin K, Kekäläinen J. Cumulated gain-based evaluation of IR techniques. ACM Trans Inf Syst. 2002;20(4):422–446. Available from: <http://portal.acm.org/citation.cfm?doid=582415.582418>.
- [149] Gey FC, Kando N, Peters C. Cross-Language Information Retrieval: the way ahead. Information Processing and Management. 2005 Dec;41(3):415–431.
- [150] Ogden W, Cowie J, Davis M, Ludovik E, Molina-salgado H, Shin H. Getting Information from Documents You Cannot Read: An Interactive Cross-Language Text Retrieval and Summarization System; 1999. .
- [151] Ogden WC, Davis MW. Improving Cross-Language Text Retrieval with Human Interactions. vol. 3; 2000. p. 3044.
- [152] He D, Wang J, Oard D, Nossal M. Comparing user-assisted and automatic query translation. Advances in cross-language information retrieval. 2003;p. 400–415.

- [153] Peters C, Deselaers T, Ferro N, Gonzalo J, Jones GJF, Kurimo M, et al., editors. Evaluating Systems for Multilingual and Multimodal Information Access: 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008. vol. 5706 of Lecture Notes in Computer Science. Springer; 2009.
- [154] Gonzalo CPKJ J. Overview of iCLEF 2008: search log analysis for Multilingual Image Retrieval. In: Borri NA F, Peters C, editors. CLEF 2008 Workshop Notes; 2008. .
- [155] Peinado V, Artiles J, Gonzalo J, Barker E, López-Ostenero F. FlickLing: A multilingual search interface for Flickr. CLEF 2008 Proceedings. 2008 Aug;p. 15.
- [156] Vundavalli S. Mining the Behaviour of users in a Multilingual Information Access Task. CLEF 2008 Proceedings. 2008 Aug;p. 1–5. Foo!
- [157] Tanase DI, Kapetanios E. Evaluating the Impact of Personal Dictionaries for Cross-Language Information Retrieval of Socially Annotated Images. CLEF 2008 Proceedings. 2008 Aug;p. 1–7.
- [158] Hansen P, Petrelli D, Karlgren J, Beaulieu M, Sanderson M. User-centered interface design for cross-language information retrieval. SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. 2002 Aug;Available from: <http://portal.acm.org/citation.cfm?id=564376.564455>.
- [159] Petrelli D. On the role of user-centred evaluation in the advancement of interactive information retrieval. Information Processing & Management. 2008 Oct;44(1):22–38.
- [160] Petrelli D, Beaulieu M, Sanderson M, Demetriou G, Herring P, Hansen P. Observing users, designing clarity: A case study on the user-centered de-

- sign of a cross-language information retrieval system. *J Am Soc Inf Sci*. 2004 Aug;55(10):923–934.
- [161] Petrelli D, Hansen P, Beaulieu M, Sanderson M. User Requirement Elicitation for Cross-Language Information Retrieval. *The New Review of Information Behaviour Research*. 2002 Aug;3.
- [162] Hansen P, Karlgren J. Effects of foreign language and task scenario on relevance assessment. *Journal of Documentation*. 2005;61(5):623–639.
- [163] Ogden W, Zacharski R, An S, Ishikawa Y. User choice as an evaluation metric for web translation services in cross language instant messaging applications. *MT Summit XII: The twelfth Machine Translation Summit*. 2009 Apr; Available from: <http://www.mt-archive.info/MTS-2009-TOC.htm>.
- [164] Ogden WC. A Task-based Evaluation Method for Embedded Machine Translation in Instant Messaging Systems. In: *Advanced Decision Architecture for the Warfighter: Foundation and Technology*. Alion Science and Technology Corp.; 2009. .
- [165] Aust R, Kelley MJ, Roby W. The Use of Hyper-Reference and Conventional Dictionaries. *Educational Technology Research & Development*. 1993 May;41(4):63–73.
- [166] Davis JN. Computers and L2 Reading: Student Performance, Student Attitudes. *Foreign Language Annals*. 1997 Oct;30(1):58–72.
- [167] Sakar A, Ercetin G. Effectiveness of hypermedia annotations for foreign language reading. *Journal of Computer Assisted Learning*. 2005 Jan;21:28–38.

- [168] Akyel A, Ercetin G. Hypermedia Reading Strategies Employed by Advanced Learners of English. *System: An International Journal of Educational Technology and Applied Linguistics*. 2009 Mar;37(1):136–152.
- [169] Chun D. L2 Reading on the Web: Strategies for Accessing Information in Hypermedia. *Computer Assisted Language Learning*. 2001;14(5):367–403.
- [170] Roseblat G, Gemoets D, Browne AC, Tse T. Machine translation-supported cross-language information retrieval for a consumer health resource. *AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium*. 2003 Jan;p. 564–8. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/14728236?dopt=abstract>.
- [171] Kit C, Wong TM. Comparative Evaluation of Online Machine Translation Systems with Legal Texts. *Law Library Journal*. 2008 Jan;100(2):299–321. Available from: http://heinonlinebackup.com/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/llj100§ion=28.
- [172] Rogati M, Yang Y. Resource selection for domain-specific cross-lingual IR. *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. 2004 Jul; Available from: <http://portal.acm.org/citation.cfm?id=1008992.1009021>.
- [173] Sharif I, Tse J. Accuracy of Computer-Generated, Spanish-Language Medicine Labels. *Pediatrics*. 2010 Apr;125:960–965. Available from: <http://pediatrics.aappublications.org/cgi/reprint/peds.2009-2530v1>.
- [174] Bedrick SD, Mauro A. A multi-lingual web service for drug side-effect data. *AMIA Annual Symposium proceedings / AMIA Sym-*

- posium AMIA Symposium. 2009 Jan;2009:34–8. Available from: http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=20351818&dopt=abstractplus.
- [175] Daumke P, Markó K, Poprat M, Schulz S. Multilingual biomedical dictionary. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2005 Jan;p. 933. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/16779220?dopt=abstract>.
- [176] Lu WH, Lin RSJ, Chan YC, Chen KH. Overcoming terminology barrier using Web resources for cross-language medical information retrieval. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2006 Jan;p. 519–23. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/17238395?dopt=abstract>.
- [177] Lu WH, Lin SJ, Chan YC, Chen KH. Semi-automatic construction of the Chinese-English MeSH using Web-based term translation method. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2005 Jan;p. 475–9. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/16779085?dopt=abstract>.
- [178] Liu F, Fontelo P, Ackerman M. BabelMeSH2 and PICO Linguist2: combined language search for MEDLINE/PubMed. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2007 Jan;p. 1036. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/18694134?dopt=abstract>.
- [179] Mahmoud MA, Al-Khafaji JTJ, Al-Shorbaji N, Sara K, Al-Ubaydli M, Ghazzaoui R, et al. BabelMeSH and PICO Linguist in Arabic. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2008

- Jan;p. 944. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/18998782?dopt=abstract>.
- [180] Bland SK, Noblitt JS, Armington S, Gay G. The Naive Lexical Hypothesis: Evidence from Computer-Assisted Language Learning. *Modern Language Journal*. 1990;74(4):440–450.
- [181] Weaver W. Translation. In: Locke WN, Booth AD, editors. *Machine translation of languages: fourteen essays*. Published jointly by Technology Press of the Massachusetts Institute of Technology and Wiley, New York; 1955. .
- [182] Anonymous. Mokusatsu: One Word, Two Lessons. *NSA Technical Journal*. 1968 Aug;13(4):95–100. Available from: http://www.nsa.gov/public_info/_files/tech_journals/mokusatsu.pdf.
- [183] Lamport L; 1987 [cited October 18, 2010]. Available from: <http://research.microsoft.com/en-us/um/people/lamport/pubs/distributed-system.txt>.
- [184] W3C. Web Services Glossary;. Available from: <http://www.w3.org/TR/ws-gloss/> [cited 12/8/2010].
- [185] Romano P, Marra D, Milanesi L. Web services and workflow management for biological resources. *BMC Bioinformatics*. 2005 Dec;6 Suppl 4:S24. Available from: <http://www.biomedcentral.com/1471-2105/6/S4/S24>.
- [186] Wright A, Sittig DF. SANDS: a service-oriented architecture for clinical decision support in a National Health Information Network. *Journal of biomedical informatics*. 2008 Dec;41(6):962–81. Available from: http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WHD-4S26662-1&_user=

1072900&_rdoc=1&_fmt=&_orig=search&_sort=d&view=c&_acct=C000048262&_version=1&_urlVersion=0&_userid=1072900&md5=e59e1476288344c7956947f016b8beb1.

- [187] Ramsay J, Barbesi A, Preece J. A psychological investigation of long retrieval times on the World Wide Web. *Interacting with Computers*. 1998;10(1):77 – 86. HCI and Information Retrieval. Available from: <http://www.sciencedirect.com/science/article/B6V0D-3T0SWJR-5/2/e74918415e2ce6c2e86d03991c859833>.
- [188] Nah FFH. A study on tolerable waiting time: how long are Web users willing to wait? *Behaviour Information Technology*. 2004;23(3):153–163.
- [189] Selvidge PR, Chaparro BS, Bender GT. The world wide wait: effects of delays on user performance. *International Journal of Industrial Ergonomics*. 2002;29(1):15 – 20. Available from: <http://www.sciencedirect.com/science/article/B6V31-44JJ91S-2/2/ff7271cae0b41de7f0342e442af2fbef>.
- [190] National Library of Medicine. Entrez Programming Utilities; 2010. Available from: <http://eutils.ncbi.nlm.nih.gov/> [cited Jan. 1, 2010].
- [191] Goto N, Prins P, Nakao M, Bonnal R, Aerts J, Katayama T. BioRuby: bioinformatics software for the Ruby programming language. *Bioinformatics*. 2010 Oct;26(20):2617–9. Available from: <http://bioinformatics.oxfordjournals.org/content/26/20/2617.long>.
- [192] Google, Inc . Google Translate; 2010. Available from: <http://translate.google.com> [cited Jan. 1, 2010].
- [193] Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology

and workflow process for providing translational research informatics support. *J Biomed Inform.* 2009 Apr;42(2):377–81. Available from: http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WHD-4TJTX88-1&_user=1072900&_coverDate=04%252F30%252F2009&_rdoc=1&_fmt=high&_orig=search&_origin=search&_sort=d&_docanchor=&view=c&_acct=C000048262&_version=1&_urlVersion=0&_userid=1072900&md5=289e5db89132748d556e3e7f06677324&searchtype=a.

- [194] Manning CD, Raghavan P, Schütze H. Introduction to information retrieval. New York: Cambridge University Press; 2008. Available from: <http://www.loc.gov/catdir/enhancements/fy0810/2008001257-b.html>.
- [195] Knuth DE. The art of computer programming. 3rd ed. Reading, Mass.: Addison-Wesley; 1997.
- [196] Peinado V, Gonzalo J, Artiles J, López-Ostenero F. UNED at iCLEF 2008: Analysis of a large log of multilingual image searches in Flickr. 2008 Aug;p. 14.
- [197] Ross S. Self-assessment in second language testing: a meta-analysis and analysis of experiential factors. *Language Testing.* 1998 Jan;15(1):1–20. Available from: <http://ltj.sagepub.com/cgi/content/abstract/15/1/1>.
- [198] Ergonomic requirements for office work with visual display terminals (VDTs) – Part 11 : Guidance on usability. ISO, Geneva, Switzerland; 1998. ISO 9241-11.

- [199] Shneiderman B, Plaisant C. Designing the user interface: strategies for effective human-computer interaction. 4th ed. Boston: Pearson/Addison Wesley; 2004.
- [200] Badjatia N, Carney N, Crocco TJ, Fallat ME, Hennes HMA, Jagoda AS, et al. Guidelines for prehospital management of traumatic brain injury 2nd edition. *Prehosp Emerg Care*. 2008 Jan;12 Suppl 1:S1-52.
- [201] Rondina C, Videtta W, Petroni G, Lujan S, Schoon P, Mori LB, et al. Mortality and morbidity from moderate to severe traumatic brain injury in Argentina. *J Head Trauma Rehabil*. 2005 Jan;20(4):368-76. Available from: <http://meta.wkhealth.com/pt/pt-core/template-journal/lwwgateway/media/landingpage.htm?issn=0885-9701&volume=20&issue=4&spage=368>.
- [202] National Library of Medicine. Summary of Enhancements for Clinical Queries for MEDLINE for Studies. *NLM Technical Bulletin*. 2004;(336).
- [203] National Library of Medicine. PubMed Clinical Queries Table;. Available from: <http://www.ncbi.nlm.nih.gov/entrez/query/static/clinicaltable.html>.
- [204] Fleiss JL, Levin BA, Paik MC. Statistical methods for rates and proportions. 3rd ed. Hoboken, N.J.: J. Wiley; 2003. Available from: <http://www.loc.gov/catdir/bios/wiley042/2002191005.html>.
- [205] Van Rijsbergen CJ. Information retrieval. 2nd ed. London: Butterworths; 1979.
- [206] Karlgren J, Hansen P. SICS at iCLEF 2002: cross-language relevance assessment and task context. *Advances in cross-language information retrieval*:

- third workshop of the cross-language evaluation forum, CLEF 2002. 2003;p. 383–391.
- [207] Sperling G, Doshier BA. Strategy and Optimization in Human Information Processing. In: Boff KR, Kaufman L, Thomas JP, editors. Handbook of perception and human performance. vol. 1. New York: Wiley; 1986. .
- [208] Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Quality of Reporting of Meta-analyses. Lancet. 1999 Nov;354(9193):1896–900. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0140673699041495>.
- [209] R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2010.
- [210] Rosner B. Fundamentals of biostatistics. 6th ed. Belmont, CA: Thomson-Brooks/Cole; 2006. Available from: <http://www.loc.gov/catdir/enhancements/fy1103/2004117046-b.html>.
- [211] Zeng QS, Li CF, Zhang K, Liu H, Kang XS, Zhen JH. Multivoxel 3D proton MR spectroscopy in the distinction of recurrent glioma from radiation injury. J Neurooncol. 2007 Aug;84(1):63–9.
- [212] Coats JS, Freeberg A, Pajela EG, Obenaus A, Ashwal S. Meta-analysis of apparent diffusion coefficients in the newborn brain. Pediatr Neurol. 2009 Oct;41(4):263–74.
- [213] Miñambres E, Ballesteros MA, Mayorga M, Marin MJ, Muñoz P, Figols J, et al. Cerebral apoptosis in severe traumatic brain injury patients: an in vitro, in vivo, and postmortem study. J Neurotrauma. 2008 Jun;25(6):581–91.

- [214] Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977 Mar;33(1):159-74. Available from: http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=843571&dopt=abstractplus.