Linguistic Resource Creation for Research and Technology Development: A Recent Experiment

STEPHANIE STRASSEL, MIKE MAXWELL, CHRISTOPHER CIERI
University of Pennsylvania Linguistic Data Consortium

Advances in statistical machine learning encourage language-independent approaches to linguistic technology development. Experiments in "porting" technologies to handle new natural languages have revealed a great potential for multilingual computing, but also a frustrating lack of linguistic resources for most languages. Recent efforts to address the lack of available resources have focused either on intensive resource development for a small number of languages or development of technologies for rapid porting. The Linguistic Data Consortium recently participated in an experiment falling primarily under the first approach, the *surprise language exercise*. This article describes linguistic resource creation within this context, including the overall methodology for surveying and collecting language resources, as well as details of the resources developed during the exercise. The article concludes with discussion of a new approach to solving the problem of limited linguistic resources, one that has recently proven effective in identifying core linguistic resources for less common studied languages.

Categories and Subject Descriptors: I.2.7 [**Artificial Intelligence**]: Natural Language Processing

General Terms: Design, Experimentation, Languages

Additional Key Words and Phrases: Machine translation, language parsing and understanding, text analysis, linguistic resources, Hindi, Cebuano, translingual information access technology, machine translation, cross-language information retrieval, information extraction, summarization

This research was supported in part by funding from the DARPA TIDES Program.
Authors' address: Linguistic Data Consortium. 3600 Market Street, Suite 810, Philadelphia PA 19104-2653

1. INTRODUCTION

Recent applications of statistical machine-learning algorithms to linguistic technologies have produced systems that are capable of both learning and improving their performance when exposed to sufficient quantities of appropriately labeled training data. Experiments in porting such technologies have revealed both the general potential for intensively multilingual computing and the specific cases in which simplifying assumptions and implementation decisions block true generality [Psutka et al. 2003; Byrne et al. 1999]. It has become clear, however, that the major impediment to creating linguistic technologies in more than a handful of the most common languages is the dearth of training data [Furui 2001; Kirchoff et al. 2002].

Attempts to address this lack of available resources have taken one of two approaches: (1) intensive effort on a small number of new languages [Cieri and Liberman 2002] and (2) development of technologies that may be rapidly ported to new languages [Al-Onaizan et al. 1999]. In the sections that follow we describe recent and ongoing work at the University of Pennsylvania's Linguistic Data Consortium as part of the *surprise language exercise,* an experiment in rapid linguistic resource and technology development largely falling under the first approach. We conclude with a discussion of yet a third approach to the problem of resource scarcity, motivated in part by our experiences in the surprise language exercise. We describe our evolving methodology and outline a plan of action, already in its beginning stages, that promises to provide core resources for a large number of critical languages. We intend this article as a call for international collaboration among resource providers and technology developers to resolve the language resource availability problem.

2. DEFINITIONS

One should note at the outset that the terms "core language resources" and "critical lesser-studied languages" are variably defined among the scholars who use them.

Consider first the issue of critical but lesser-studied languages. According the *Ethnologue* [Grimes 2003], there are nearly 7000 languages spoken in the world today. The thought of creating resources for all of them boggles the imagination and represses further discussion or planning. Here we propose to focus on a manageable subset, those that are the native languages of at least one million people. This reduces our scope to some 300 languages. Nearly 80% of the world’s inhabitants speak one of these languages natively. Figure 1 plots the cumulative distribution of the world’s inhabitants by their native languages. This graph shows that most of the world's inhabitants are native speakers of the 320 most common languages. It is clear that creating resources for this set of languages provides the biggest benefit for the effort.

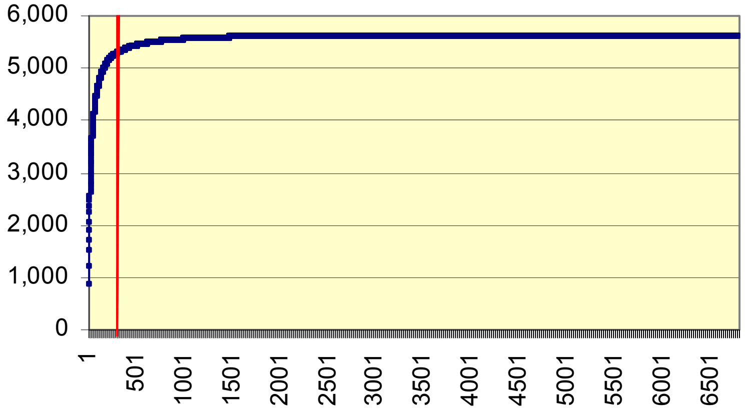


Fig. 1. Cumulative distribution of the world's inhabitants (y-axis) by native language (x-axis). The 320 most common native languages cover 80% of the world's inhabitants.

As for core language resources, we define these as the resources necessary for translingual information access technologies. Such resources include texts, parallel texts, translation lexicons, entity databases plus a range of manual annotations designed to provide training material as well as benchmark test data. Clearly, this list provides for only a subset of the desirable technologies; however, we believe that these are the critical resources for translingual information access, as well as being those resources that are currently within our grasp.

3. THE SURPRISE LANGUAGE EXERCISE

The Linguistic Data Consortium recently participated, along with several other research sites, in an experiment known as the surprise language exercise. The exercise challenged sites to identify or create linguistic resources and develop working technology for a previously untargeted language within a constrained time span.

The exercise was part of the DARPA program in Translingual Information Detection, Extraction and Summarization (TIDES), which requires computer-readable resources sufficient to support translingual information processing tasks [Wayne 2002]. Although TIDES had adopted an early focus on common languages such as English, Chinese, and Arabic where ready availability of data would allow research to continue relatively unfettered, the porting of TIDES technologies to less common languages has always been a *desideratum* of the program. For the primary focus languages, TIDES has already produced a very rich set of resources, described elsewhere [Cieri and Liberman 2002; LDC 2003]. In 2003, the program began to address the need for technologies in less common languages through experiments in rapid technology porting where data collection, resource creation, and technology development take place simultaneously within a very short time period (i.e., one month).

During the surprise language exercise described below, LDC's primary role was to coordinate development and dissemination of linguistic resources for the target language. Once the desired resources had been obtained or created, technology sites put them to use in developing NLP tools for statistically-based machine translation, topic detection and tracking, cross-lingual information retrieval, information extraction, and summarization. While evaluation of this work is ongoing, preliminary results are promising.

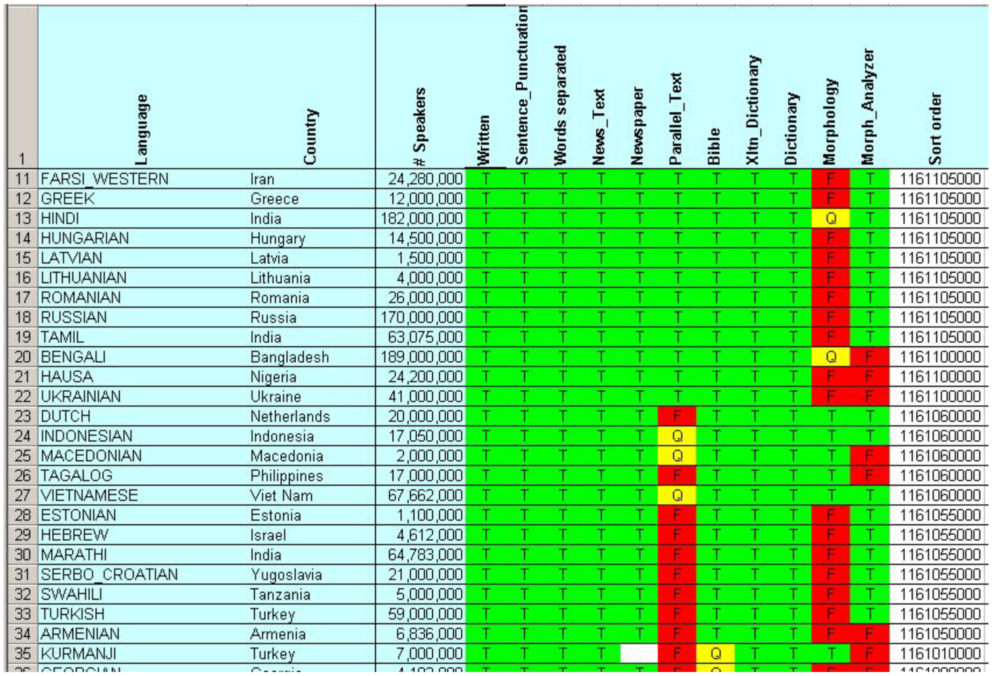
4. PREPARATION: A LANGUAGE RESOURCES SURVEY

In preparation for the surprise language exercise, LDC staff designed and began to implement a survey of language resources for the 320 most common languages. A complete description of the survey would take us beyond the scope of the present article, but the survey questions explore the structural features of a language, the demographic features of its speakers and the availability of linguistic resources, digital or otherwise, to support technology development. A linguist completes the questions of the survey in an order that allows quick scoring of languages according to their compatibility with the kinds of technology we hope to support. For example, one of the first questions is whether the language is written. There are several languages with more than one million speakers but which have no tradition of literacy, making them impractical targets for technologies that rely on large volumes of written material. LDC has completed the survey in part or wholly for over 150 languages, and plans to continue the survey for the remainder of the 320 as time and funding allow.

Figure 2 shows a sample of a summary report produced from the full survey. The categories across the top of the spreadsheet show some of the items covered by the survey that were of special importance for the choice of surprise language, including the main country where the language is spoken, number of native speakers, whether the language is written, whether the survey found news text and other resources in electronic form, whether the language has a "complex" morphology, and so on. The final column displays a numeric summary of the "true" and "false" answers to each question (and in some cases, a "questionable" answer); this is used to sort the languages by candidate status.

The actual survey report includes details for each of the categories displayed in the summary report, as well as for a number of other categories. For example, if the answer to (electronic) "News_text" is true, the detailed survey report would list URLs for the news websites or other sources of news text that we had identified.

We believe that the results of this survey will be of interest to a wide variety of users; moreover, others will be able to fill in gaps in our knowledge to further enrich the survey. While our intention is to eventually publish the survey, we have not yet determined the manner in which it will be made available.



1	Language	Country	# Speakers	Written	Sentence_Punctuation	Words separated	News_Text	Newspaper	Parallel_Text	Bible	Xltn_Dictionary	Dictionary	Morphology	Morph_Analyzer	Sort order
11	FARSI_WESTERN	Iran	24,280,000	T	T	T	T	T	T	T	T	T	F	T	1161105000
12	GREEK	Greece	12,000,000	T	T	T	T	T	T	T	T	T	F	T	1161105000
13	HINDI	India	182,000,000	T	T	T	T	T	T	T	T	T	Q	T	1161105000
14	HUNGARIAN	Hungary	14,500,000	T	T	T	T	T	T	T	T	T	F	T	1161105000
15	LATVIAN	Latvia	1,500,000	T	T	T	T	T	T	T	T	T	F	T	1161105000
16	LITHUANIAN	Lithuania	4,000,000	T	T	T	T	T	T	T	T	T	F	T	1161105000
17	ROMANIAN	Romania	26,000,000	T	T	T	T	T	T	T	T	T	F	T	1161105000
18	RUSSIAN	Russia	170,000,000	T	T	T	T	T	T	T	T	T	F	T	1161105000
19	TAMIL	India	63,075,000	T	T	T	T	T	T	T	T	T	F	T	1161105000
20	BENGALI	Bangladesh	189,000,000	T	T	T	T	T	T	T	T	T	Q	F	1161100000
21	HAUSA	Nigeria	24,200,000	T	T	T	T	T	T	T	T	T	F	F	1161100000
22	UKRAINIAN	Ukraine	41,000,000	T	T	T	T	T	T	T	T	T	F	F	1161100000
23	DUTCH	Netherlands	20,000,000	T	T	T	T	T	F	T	T	T	T	T	1161060000
24	INDONESIAN	Indonesia	17,050,000	T	T	T	T	T	Q	T	T	T	T	T	1161060000
25	MACEDONIAN	Macedonia	2,000,000	T	T	T	T	T	Q	T	T	T	T	F	1161060000
26	TAGALOG	Philippines	17,000,000	T	T	T	T	T	F	T	T	T	T	F	1161060000
27	VIETNAMESE	Viet Nam	67,662,000	T	T	T	T	T	Q	T	T	T	T	T	1161060000
28	ESTONIAN	Estonia	1,100,000	T	T	T	T	T	F	T	T	T	F	T	1161055000
29	HEBREW	Israel	4,612,000	T	T	T	T	T	F	T	T	T	F	T	1161055000
30	MARATHI	India	64,783,000	T	T	T	T	T	F	T	T	T	F	T	1161055000
31	SERBO_CROATIAN	Yugoslavia	21,000,000	T	T	T	T	T	F	T	T	T	F	T	1161055000
32	SWAHILI	Tanzania	5,000,000	T	T	T	T	T	F	T	T	T	F	T	1161055000
33	TURKISH	Turkey	59,000,000	T	T	T	T	T	F	T	T	T	F	T	1161055000
34	ARMENIAN	Armenia	6,836,000	T	T	T	T	T	F	T	T	T	F	F	1161050000
35	KURMANJI	Turkey	7,000,000	T	T	T	T		F	Q	T	T	T	F	1161010000

Fig. 2. Sample from survey of language resources.

Although work on the survey began before the one month allocated for the surprise language exercise, we argued that such a head start was in fact appropriate because (1) it was conducted in a general way without knowing which language would be the specific target of the experiment; (2) it allowed TIDES sponsors to select from the set of languages where rapid porting was an actual possibility; and (3) it changed the terrain both fundamentally and permanently for those who would port linguistic technologies to less common languages. In other words, once the survey was made available, the task of rapid porting became easier for a large number of languages. It was this realization, along with the success of the experiment described in the section below, that impels us toward the more ambitious proposal presented in the article's final section.

5. THE DRY-RUN: AN EXPERIMENT IN RESOURCE DISCOVERY

In March 2003, a surprise language dry-run was organized by LDC to assess the feasibility of the full-scale experiment and to answer basic questions about how best to administer a large-scale, collaborative, rapid resource, and technology-development exercise. Neither LDC nor any other participating sites knew in advance what language would be selected by TIDES sponsors for the dry-run. On March 5, participants were notified that the target was Cebuano, a language of the Philippines. Prior searches for computer-readable data on this language had turned up only a bible and one small news text archive. (As it turned out, this news archive contained fewer than 10,000 Cebuano words.) In addition, we knew of several printed dictionaries and grammars.

Within eight hours of the beginning of the exercise, a team of eight linguists and programmers at LDC had discovered 250,000 words of news texts in Cebuano, several other small monolingual and bilingual Cebuano texts, and no fewer than four computer-readable lexicons, one of which turned out to have on the order of 24,000 entries (lexemes). Other sites working on the exercise identified resources as well. There was a good deal of overlap among what different sites discovered, giving us some confidence that we had located a reasonably complete set of resources.

The disparity between what we were able to find before versus during the exercise is attributable in part to the greater effort during the exercise: a few person-hours before, eight hours times eight participants during. But perhaps more important was the search methodology. Prior to the exercise, we had done searches for the word “Cebuano” in combination with other English words, such as “lexicon,” “dictionary,” or “news.” But this missed some resources that were labeled with alternative names for Cebuano (Bisayan and Visayan), as well as resources that were not labeled as dictionaries, etc.

During the exercise, we employed a different method, suggested by Mark Liberman (see also Ghani et al. [2001]). Once we had found a handful of pages in Cebuano, we did a count of word forms. We fed two of the most common word forms (the words for “this” and “that”) back into search engines, and this quickly led to more discoveries. This technique, using these and other common words, led to most of our discoveries. We also used lists of words and tetragrams as queries to electronic lexicons of Cebuano to determine the extent of their coverage.

Since the Cebuano dry-run, we have experimented with the technique of searching for resources using seed words in new languages. Preliminary experiments have been promising, even with languages having extensive inflectional morphology. For instance, Tzeltal (a Mayan language of Mexico), Swahili (east Africa), and Shuar (a Jivaroan language of Ecuador) all have substantial inflectional morphology, including both prefixes and suffixes. Nevertheless, searches with a few common nouns return numerous hits, most of which are indeed texts in the target languages.

Seed terms can be extracted from an initial set of texts, as we did with Cebuano and Swahili; or they can come from dictionaries, as was the case for Tzeltal and Shuar. In the case of inflectionally rich languages, when the word forms are extracted from texts, they will obviously be inflected. Dictionary citation forms for most languages are normally inflected forms as well, although dictionary writers commonly strive to use the least inflected form of a word. But for some languages, dictionaries traditionally use citation forms that are bare roots which can never appear in texts in that form. Using this search technique with such languages would require consulting a grammar to create inflected forms to be fed into search engines.

Word length is also a consideration. Words that are just two or three characters long may turn up too frequently in other languages to be of use. Where there are closely related languages, even longer words may give rise to spurious hits in the related languages. Ideally, we would have lexicons of a 100 or more languages, and use these lexicons to eliminate candidate search terms that appear in other languages. (Small lexicons would probably serve this purpose better than large lexicons, since homographs across languages are only problematic if they are common in the other languages.)

Another issue is the fact that for minority languages, the writing system may have undergone recent changes. An older dictionary may contain words that are now spelled differently, and its use will therefore result in a lack of hits. This was true of some printed dictionaries of Cebuano; fortunately, the spelling changes required were mechanical.

Encodings present additional issues. Typing in search terms in an encoding requires choosing a keyboard and encoding, and some encodings may not be supported. Much simpler is copying and pasting seed terms from a web page in the appropriate encoding.

However, some languages use multiple encodings. For instance, several languages of eastern Europe have more than one commonly used encoding. A worst case is Amharic (of Ethiopia), with over 70 encodings (see the LibEth project at http://libeth.sourceforge.net); several of which are commonly used on web pages. In order to get a sufficiently broadly based corpus, it may therefore be necessary to enter search terms in multiple encodings.

One additional issue we faced during the Cebuano dry-run was that of language identification. Cebuano is related to a number of other Philippine languages (and more distantly to other Malayo-Polynesian languages), and it can therefore be difficult for nonspeakers to tell whether texts are actually in Cebuano. We addressed this problem by looking up a variety of words from the texts in question in both printed and computer-readable Cebuano dictionaries. This only works for words with no inflectional affixes, so the recall of this method is limited by inflectional morphology. Cebuano has minimal morphology (verbs are inflected with prefixes, infixes, and suffixes, but nouns are for the most part uninflected); hence we were reasonably certain that our texts were in Cebuano. Lingering doubts having to do with languages that are very closely related to Cebuano were removed by having Cebuano speakers check the texts.

In sum, during this short exercise, we were able to locate a surprising number of resources in a short period of time, giving us confidence that for a full-scale exercise we would be able to find sufficient resources in any language that appeared practical based on our preliminary survey.

6. THE TEST: THE SURPRISE LANGUAGE EXERCISE

A few months after the work described above, the full surprise language exercise took place. Unlike the dry-run, which was designed primarily to evaluate the process for rapidly locating and disseminating linguistic resources, the full exercise would put those resources to use in the development of various natural language processing technologies. On June 2, 2003, participants learned that the surprise language was Hindi. Both the process and the results of the full exercise were significantly different from the dry-run. For one, the amount of text written in Hindi available on the web is orders of magnitude greater than for Cebuano. Thus we were not faced with the difficulty of finding Hindi text, but rather with processing vast quantities of it.

A second difference—and one which loomed ever larger in our minds during the course of the experiment—arose from the fact that while Cebuano is written with a Latin alphabet, and can therefore be encoded with the ASCII character set, Hindi uses the Devanagari writing system. Indian computer scientists have therefore developed an 8-bit character encoding known as ISCII (see http://brahmi.sourceforge.net/docs/iscii91.pdf for a draft standard), which reportedly forms the basis of the Unicode implementation for Hindi. In addition, there are several different Romanization standards.

Unfortunately, we encountered no web site from which to harvest text that actually used ISCII, and neither of the news sites that used Unicode (Voice of America and the BBC) was based in India. Instead, virtually every Hindi news site had its own more or less proprietary 8-bit font, and each font used its own unique encoding. Indeed, several web sites used more than one font and/or encoding; the India parliament requires downloading five different fonts, although some of these appear to use the same encoding.

In order to develop NLP tools that would work with text from different web sites, we were forced to convert all the text to a standard encoding, for which we chose Unicode (UTF-8).

Written Hindi has around 50 consonants and vowels, with no upper/lower case distinctions. This would easily fit into a 7-bit character set (or into the upper 128 code points of an 8-bit character set). However, there are variant forms of many consonants used when these appear in consonant clusters, as well as variant forms of vowels. The ISCII character set assumes intelligent font rendering, so that only a single form of each consonant or vowel needs to be encoded. But most designers of proprietary fonts have encoded variant character forms, electing instead to put the intelligence into keyboard drivers. The result is that not only are the code points different for each font, the set of characters which are actually encoded are to some extent different—rendering the encoding conversion process nontrivial.

An analogy to this problem in Roman character encodings is accented characters, which under some conventions are treated as unitary characters, while other conventions treat them as a base character plus diacritical marks. Choices between unitary and multigraph representation are prevalent in Hindi, and different alternatives are commonly used in different encodings.

Likewise, some English typesetting conventions provide special treatment for certain character sequences, such as “fl”: the shape of the individual characters may be slightly modified in such ligatures. Ligature-like characters are abundant in Hindi, and encodings often provide hard-coded ligature forms; again, the decisions as to which ligature forms to hard-code are often made differently for different Hindi encodings.

Debugging character-set conversion also proved difficult. We often found that a converter that appeared to work on a small text sample failed to completely convert larger texts, leaving the result peppered with errors, both visible and covert. (By “covert” errors we mean errors that, while not visible in displayed text, result in differences in the underlying sequence of code points, and would therefore affect, e.g., dictionary lookup.) Some of these errors were due to bugs in our rapidly developed converter, while others were due to nonstandard characters or character sequences in the text (e.g., in loan words), or simply typos (which were surprisingly frequent in some texts).

In sum, character-set conversion turned out to be a much greater problem than we had anticipated. Despite these difficulties, a significant number of resources were identified, converted, and further processed by LDC with much help from the other sites participating in the exercise.

While the resource discovery period for Cebuano lasted just a few hours, the same process extended into the final days of the Hindi exercise. This heavily collaborative effort garnered at least 13 lexicons (both general and domain-specific), resulting in nearly 30,000 unique lexical entries; 17 sources of monolingual text; over 30 sources of bilingual text, including the bible and other literature but also several news and government websites plus several bilingual corporate and technology websites. Participants also found a general morphological parser and a number of entity lists, including telephone directories, a geographical place name list, and government voter and personnel lists.

During the Cebuano exercise, we had experimented with the use of a “wiki” (publicly editable) web site for purposes of resource dissemination, and with a blog for rapid communication, but neither technique seemed to work terribly well. In particular, we encountered problems with conflicting multiple edits, presumably caused by the very rapid posting and the need for continual updating of information. As a result, it became necessary for each site to have its own web page on the wiki, thereby eliminating the supposed advantage of *collaborative* editing. Moreover, with multiple pages on the wiki, a visitor in search of a particular resource had to consult numerous pages, making the sharing of resources somewhat cumbersome.

Accordingly, for the Hindi exercise we chose a centrally located and edited repository for all shared data, which worked in the following way.

When a site identified a resource, it was announced through an email listserv to all participants; the emails could also be accessed from a web archive. URLs of all "found" resources were posted on a web page whose URL was made available to all participants.

Particularly interesting items on the found resources list were then selected by one of the participating sites for download and further processing. In the case of text resources, processing might involve identifying encoding, transliteration into a standard encoding (in the case of Hindi), stripping HTML tags, and tokenization. In some cases multiple versions of resources might be provided, e.g., a version of a text with improved encoding conversion, or a lexicon consisting of merged lexical entries from a number of found lexicons.

In addition to processing found resources, some sites created new resources from scratch, such as morphological stemmers or encoding converters.

Both processed and created resources were treated in the same way. A site wishing to provide such a resource had two choices: it could either announce it in the email list and say that the file would be distributed by LDC, or it could announce it but make the file available at its own web site. The latter was a faster way to make available especially important resources, as LDC sometimes found itself a couple days behind in processing the resources found during the Hindi exercise; but it had the disadvantage of being more difficult for potential users to find and download. In either case, the file was (eventually) made available from a second, password-protected web page at LDC, providing a central download site.

The process of making submitted resources available from the LDC website was essentially a matter of validating the resource, ensuring that its contents and format were documented, and entering the information into a database system. This resulted in something of a bottleneck, with one or two individuals at LDC (who were also performing many other critical surprise language tasks) logging each of the submitted resources. This difficulty might have been avoided by allowing remote sites to upload the resource files and enter the metadata into the database themselves, but at some cost to consistency. A simpler solution might have been to have someone at LDC whose sole task was resource validation and logging.

The reason for password protection on the processed resources was that much of the raw data harvested for the surprise language exercise was drawn from commercial data providers. While use of this data for the purposes of the exercise itself can be seen as falling within the context of fair use, this limits access to the data to those TIDES sites participating in the exercise. LDC is currently pursuing intellectual property rights negotiations with data providers in order to secure distribution rights, so that much of the data developed during surprise language can eventually be made available to a wider community of linguistic researchers, educators, and technology developers.

As mentioned earlier, we also used teleconferencing to work through issues: daily at first, then two or three times weekly. Teleconferencing turned out to be a highly effective supplement to our other forms of communication, particularly for discussing problems with the resource collection process.

7. RAPID LINGUISTIC RESOURCE CREATION

Not all required resources were immediately available on the web. For both the dry-run and the full exercise, after existing stores of data for the target language had been identified and harvested, human annotators worked to create topic-relevance judgments, manual summaries, entity-tagged texts, aligned parallel text, and a host of other resources. Moreover, human annotators were needed to create answer keys for the benchmark test data used in evaluating the surprise language technology.

For the Cebuano dry-run, LDC resource creation focused on general resources: sentence-aligned bilingual text, entity-tagged data, and morphological parsers. During the month-long Hindi exercise, LDC worked with other sites to produce not only general resources but also substantial quantities of annotated and unannotated training and test data to support the full range of TIDES technologies: information extraction, detection, summarization, and machine translation. LDC also defined the training and evaluation corpora for each task and worked with the US National Institute of Standards and Technology (NIST) to distribute this data to participating sites, enabling NIST to then evaluate system performance against stable ground-truth data labeled by human judges.

In preparation for the surprise language experiments, LDC had created basic annotation tools and streamlined annotation guidelines that would allow annotators to make rapid progress on each task with minimal training. Platform-independent, multilingual annotation tools were developed for the exercise, primarily utilizing the Annotation Graph Toolkit or AGTK [Bird and Liberman 2001]. The tools take tokenized, UTF-8 encoded text as input and save annotation records as stand-off markup. This approach also allowed annotation work to be distributed across multiple sites. Particularly in the case of Hindi, both the pre-existing annotation tools and the process for creating manually tagged data had to be substantially revised to handle the encoding issues described above. This reduced the amount of time ultimately available for creation of annotated data, which is reflected in both the quality and the overall quantity of the resources created for Hindi.

News texts labeled for topic relevance are an important resource for information retrieval and related technologies. During the Hindi exercise, LDC annotators engaged in topic development and relevance assessment to support cross-language information retrieval (CLIR) and topic detection and tracking (TDT). For both evaluation areas, LDC defined a training corpus consisting of news texts drawn from the Hindi found resources. Native Hindi-speaking annotators then scanned the corpus and selected 15 broad (theme-based) topics and 15 narrow (event-based) topics. Annotators created profiles for each of the resulting topics, consisting of a title, definition, and narrative plus a set of query terms. Each topic profile was also translated into English from the original Hindi.

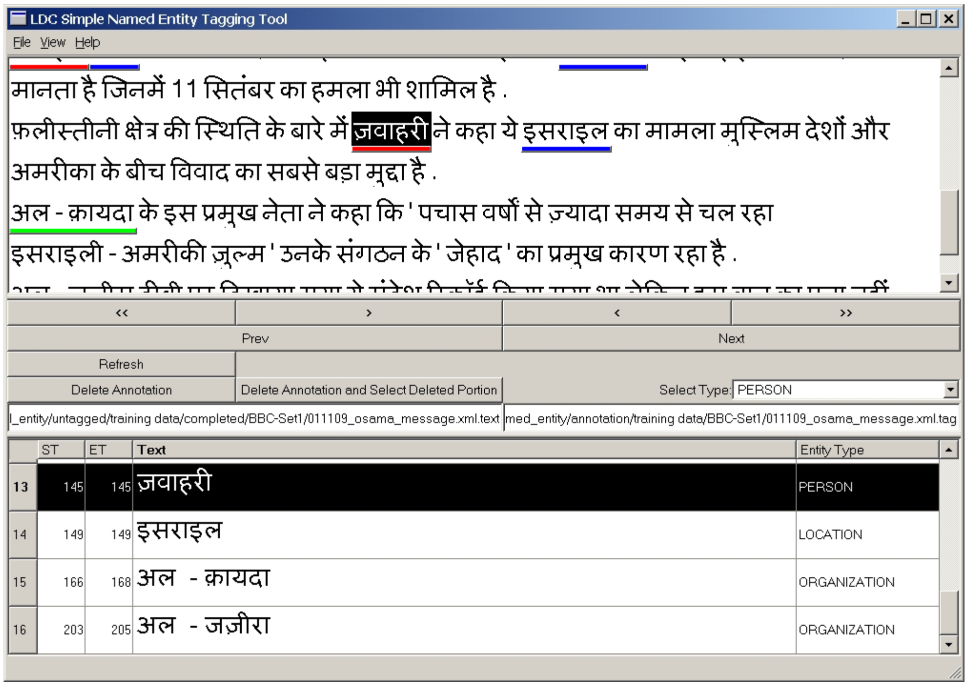
Research sites participating in CLIR were given topic profiles for the 15 broad topics to use as training data. Sites used these training topics to index the news corpus for topic relevance. Human annotators then read and labeled the news stories in the resulting relevance-ranked lists in order to establish ground-truth for each topic. During the Hindi exercise, LDC annotators labeled a total of 1710 documents for CLIR topic relevance.

For the TDT evaluation, sites were provided not with the topic profiles, but with four on-topic training documents for each of the 15 event-based topics. Systems were then required to detect all other on-topic documents in the Hindi corpus. In addition, 11 of the 15 topics were selected for cross-language detection in English. LDC annotators worked with annotators at the University of Massachusetts Amherst to complete topic development and relevance assessment of the sites' submissions.

Topic-relevance annotation was completed using LDC's existing topic-tagging toolkit developed previously for TDT, TREC, and related projects, and customized for the surprise language exercise to handle Hindi data.

Although web pages can frequently be mined for lists of certain kinds of entities (names of government officials and place names, for example), texts in which named entities have been tagged are virtually unobtainable on the web. This kind of training data is required for information-extraction technology development, thus necessitating manually annotated training data.

The named entity task for surprise language utilized a subset of the Message Understanding Conference (MUC) named entity annotation guidelines [Chincor 1997], excluding temporal expressions and number expressions from annotation. Annotators focused instead on three named entity types: *organizations*, consisting of named corporate, governmental, or other organizational entities; *persons*, consisting of named persons or families; and *locations*, which can be politically or geographically defined (e.g., cities, countries, mountain ranges, bodies of water). During the surprise language exercise, annotators at BBN, NYU, and LDC created over 430,000 words of named entity training data, including some data that was annotated twice to establish interannotator agreement rates. Annotation was performed using the AGTK named entity tagging tool that LDC had developed for the exercise. The tool allowed annotators to swipe over a region of text, and then use the mouse to select the appropriate entity type from a pull-down menu. Annotated text is displayed with color-coded underlining.


LDC Simple Named Entity Tagging Tool
File View Help
<<
>
<
>>
Prev
Next
Refresh
Delete Annotation
Delete Annotation and Select Deleted Portion

मानता है जिनमें 11 सितंबर का हमला भी शामिल है .
फ़लीस्तीनी क्षेत्र की स्थिति के बारे में ज़वाहरी ने कहा ये इसराइल का मामला मुस्लिम देशों और
अमरीका के बीच विवाद का सबसे बड़ा मुद्दा है .
अल - क़ायदा के इस प्रमुख नेता ने कहा कि ' पचास वर्षों से ज़्यादा समय से चल रहा
इसराइली - अमरीकी ज़ुल्म ' उनके संगठन के ' जेहाद ' का प्रमुख कारण रहा है .

Select Type: PERSON

l_entity/untagged/training data/completed/BBC-Set1/011109_osama_message.xml.text
med_entity/annotation/training data/BBC-Set1/011109_osama_message.xml.tag

	ST	ET	Text	Entity Type
13	145	145	ज़वाहरी	PERSON
14	149	149	इसराइल	LOCATION
15	166	168	अल - क़ायदा	ORGANIZATION
16	203	205	अल - जज़ीरा	ORGANIZATION

Fig. 3. Simple named entity annotation tool.

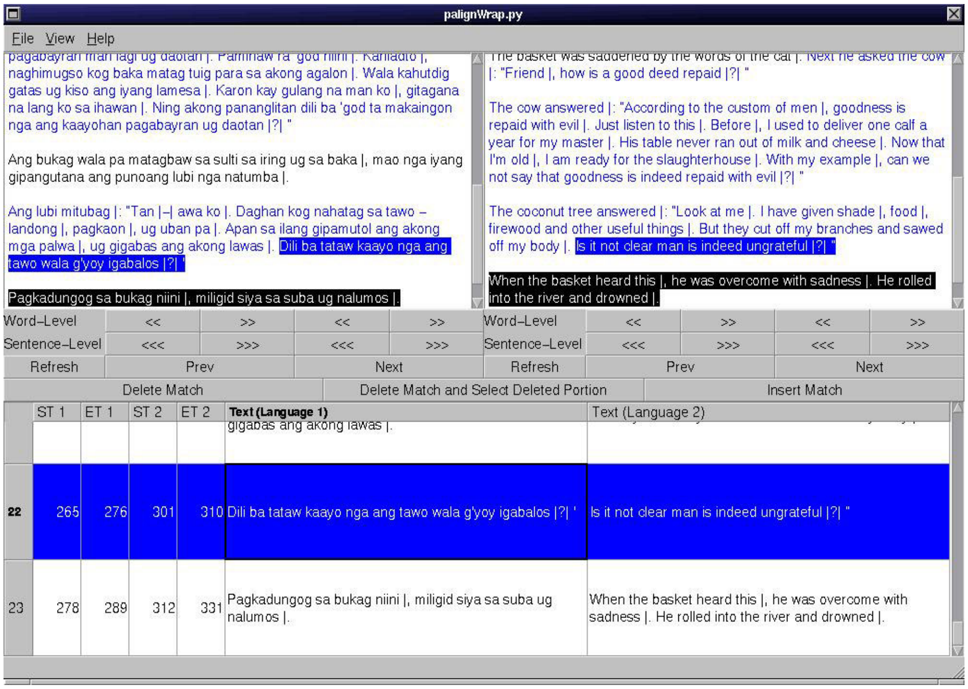
In addition to the training data described above, LDC also defined an evaluation corpus of 25 Hindi news documents. These documents were subject to additional processing and manual validation to ensure appropriate news content and consistent encoding. Despite this extra attention some minor encoding anomalies remained; this was an unfortunate side effect of the fact that all processing and manual validation of the test data had to be completed in just a few hours at the very end of the exercise through intensive collaboration between non-Hindi programmers and non-programmer Hindi speakers.

The resulting Hindi test corpus was annotated in multiple ways to provide ground-truth data for several evaluations. Annotators tagged the data for named entities for the extraction evaluation, and four independent annotators each created 10-word summaries in English for each of the test documents to support the summarization evaluation.

Bilingual texts are another critical resource. For purposes of machine translation (MT), the best bilingual training text belongs to the same genre as the text that the MT programs are expected to translate—in the case of the surprise language exercise, news text. We found some bilingual news text for Cebuano, but not as much as we needed. The only way to obtain the needed bilingual text was therefore to create it, and for this purpose we hired a number of translation agencies. At an average price of 28 cents per word, this is not an inexpensive operation, but neither is it impossible. We also used these agencies to create manual translations of the Hindi test corpus in order to provide ground-truth data for the Hindi MT evaluation.

In addition to simply having bilingual text, we wanted to align those texts at the sentence level. The bible is available in Cebuano and in Hindi, and in effect constitutes parallel text aligned at the verse level. However, the style and vocabulary of bible translation is different enough from news text that it is desirable to align bilingual news text as well. Because of the way we prepared the text to be sent to translation agencies, it came back already aligned.

But we had other bilingual texts that we found on the web, and annotators aligned these using a manual alignment tool we had built using AGTK. The two-panel annotation tool displays the original document and its translation side-by-side; the tool then automatically selects the first sentence (defined by its punctuation) in the source data, along with the first sentence in the translation. If the annotator judges these sentences to be a translation pair, s/he hits a button to record that judgment. The annotation is then stored as a record in a separate file, which indexes each translation pair in terms of token offsets. The tool then automatically selects the next sentence in both the source document and the translation, and the annotator makes another judgment. If the two sentences selected by default are not translation pairs, the tool allows the annotator to add or delete words from the selection in either language, or to jump to another complete sentence selection with ease.



pagabayran man lagi ug daotan |. Paminaw ra 'god niini |. Kanhiadto |, naghimugso kog baka matag tuig para sa akong agalon |. Wala kahutdig gatas ug kiso ang iyang lamesa |. Karon kay gulang na man ko |, gitagana na lang ko sa ihawan |. Ning akong pananglitan dili ba 'god ta makaingon nga ang kaayohan pagabayran ug daotan |?| "

Ang bukag wala pa matagbaw sa sulti sa iring ug sa baka |, mao nga iyang gipangutana ang punoang lubi nga natumba |.

Ang lubi mitubag |: "Tan |–| awa ko |. Daghan kog nahatag sa tawo – landong |, pagkaon |, ug uban pa |. Apan sa ilang gipamutol ang akong mga palwa |, ug gigabas ang akong lawas |. Dili ba tataw kaayo nga ang tawo wala g'yoy igabalos |?| '

Pagkadungog sa bukag niini |, miligid siya sa suba ug nalumos |.

The basket was saddened by the words of the cat |. Next he asked the cow |: "Friend |, how is a good deed repaid |?| "

The cow answered |: "According to the custom of men |, goodness is repaid with evil |. Just listen to this |. Before |, I used to deliver one calf a year for my master |. His table never ran out of milk and cheese |. Now that I'm old |, I am ready for the slaughterhouse |. With my example |, can we not say that goodness is indeed repaid with evil |?| "

The coconut tree answered |: "Look at me |. I have given shade |, food |, firewood and other useful things |. But they cut off my branches and sawed off my body |. Is it not clear man is indeed ungrateful |?| "

When the basket heard this |, he was overcome with sadness |. He rolled into the river and drowned |.

	ST 1	ET 1	ST 2	ET 2	Text (Language 1)	Text (Language 2)
					gigabas ang akong lawas |.	
22	265	276	301	310	Dili ba tataw kaayo nga ang tawo wala g'yoy igabalos |?| '	Is it not clear man is indeed ungrateful |?| "
23	278	289	312	331	Pagkadungog sa bukag niini |, miligid siya sa suba ug nalumos |.	When the basket heard this |, he was overcome with sadness |. He rolled into the river and drowned |.

Fig. 4.Text alignment annotation tool.

Morphological parsers can be found on the web for some languages, and indeed one was located for the Hindi exercise, but we did not find one for Cebuano. However, Cebuano inflectional morphology is fairly simple. We used a grammatical description of Cebuano [Bunye 1971], together with a merged version of the Cebuano lexicons we had found, to build a morphological transducer running under the Xerox program xfst [Beesley and Karttunen 2003]. Writing the grammatical rules for the parser took just a few hours. Testing on news texts revealed a parse rate of around 60%. Many of the “failures” turned out to be English loan words that were not listed in the Cebuano lexicons, or punctuation or number tokens. Eliminating these gave a parse rate above 90%.

Some of the remaining unparsed words in our Cebuano news texts were Spanish loans. Simply adding a Spanish parser would not work, however, because these loans (unlike most of the English loans) are spelled according to Cebuano orthography. Most of the differences are mechanical (e.g. Cebuano “k” comes from Spanish “c” or “qu”). It would not be difficult to construct a transducer to convert between the two orthographies, and then use a Spanish transducer to gloss the Spanish loans with English. But we have not attempted this step.

While (near-) native speaker annotators were preferred for all surprise language annotation tasks, it may be extremely difficult to locate and hire qualified staff for some languages; even when skilled staff can be identified they may be unable to devote time to the project or may be prohibited from working due to visa restrictions. In order to minimize the need for native speaker annotators and to reduce the amount of time required by each annotator to produce high-quality resources, LDC developed annotation practices and tools to make the process maximally efficient. In some cases, non-native speakers can perform a substantial amount of initial annotation work, and this work can be checked over by native speakers. If the native orthography of the language is familiar or a standard Romanization exists, then English-speaking annotators can achieve high accuracy on both the parallel text alignment task and some parts of the named entity task. For parallel alignment, punctuation cues, cognates, names, and numbers provide cues to sentence pairs in each language.

During the Cebuano exercise this approach was used quite successfully. English-speaking annotators manually aligned a subset of the parallel text data; when a native Cebuano speaker checked over the data, very few realignments were necessary, and most of those were not corrections but rather splitting larger chunks of text into finer alignment pairs. For the named entity task, personal names and some organization and location names may be represented identically in both English and the target language, allowing non-native speakers to perform a large portion of the tagging. Because the native orthography for Hindi does not use Latin characters, however, this method could not be exploited during the full surprise language exercise; hence a large team of native Hindi speakers had to be employed to create the range of manually tagged resources required for the experiment.

In general, annotation projects demand detailed, well-tested guidelines and customized annotation tools, careful hiring decisions followed by long periods of annotator training, and extensive quality assurance processes, all of which require substantial time and effort to implement. These processes had to be adjusted significantly to meet the special demands of the surprise language exercise. While typical annotation guidelines are quite extensive and aim to provide an exhaustive set of rules for handling rare or exceptional cases (as well as covering typical cases), the guidelines developed for the surprise language were more limited and focused only on the most common types of constructions. Team leaders had a deeper understanding of the full guidelines, and when annotators encountered a construction they did not know how to classify, the team leader could provide detailed instruction. This allowed the annotators to focus on creating data rather than learning guidelines that may never be applied to the current task. This approach was essential given both the time constraints imposed by the exercise and the potentially limited pool of native speakers, let alone native-speaking linguists or language experts, available to act as annotators for a given language.

Instead of highly customized annotation tools, for the surprise language exercise, we developed a basic suite of multilingual, platform-independent tools that could be reconfigured to meet the demands of a particular language or task. Modifications were also made to LDC's staffing procedures so that native speakers could be identified, interviewed, hired, and trained as annotators within hours or days rather than weeks.

Such changes were necessary to allow for rapid resource creation under the surprise language context.

However, these divergences from LDC's normal practices came at a cost. Without knowing in advance what the "surprise" language would be, annotation guidelines were necessarily general (and by default, English-centric). For instance, some necessary changes to the Hindi named entity guidelines became apparent only in the final week of the exercise; only then had annotators seen enough data and learned enough about the task to fully understand why some general-purpose rules were not well suited to Hindi. Also, with quick hiring and limited training, annotation quality suffered. Regular quality assurance measures like second passing, dual annotation, and discrepancy resolution had to be skipped in order to meet the aggressive surprise language deadlines. In some cases, quick updates to the annotation tools to make them display Hindi text properly introduced new bugs that had to be fixed before annotation could proceed, resulting in frustrated annotators, panicked managers, and anxious researchers (not to mention fed-up programmers!). Ultimately, it proved possible to collect or create the linguistic resources needed to enable technology development and evaluation in the context of the surprise language exercise, but not without impacting both resource quantity and quality.

As in the case of the language survey, we believe that the annotation tools developed for the surprise language exercise will be of interest to those who wish to create comparable resources for other languages. While the tools are currently optimized to work within LDC's local operating environment and within the surprise language context, we intend to make the toolkit freely available, along with documentation and perhaps training courses, once we have completed further modifications and testing.

8. SUMMARY OF RESULTS

The Cebuano dry-run and the Hindi full exercise targeted two very different languages from the standpoint of resource availability. We summarize here some of the differences.

- Cebuano was (relatively) a resource-scarce language, whereas abundant resources are available for Hindi—more, in fact, than we could actually process in a short time. Some of the resources found for each are summarized in Table I (this does not include the encoding converters which were found or built for Hindi, nor text-tagged for certain other purposes, e.g., time phrases or parts of speech).
- Cebuano was written in a Roman writing system, whereas Hindi is written in a nonRoman system. One implication of this is that it was much easier for non-native speakers to work with and even annotate Cebuano text than Hindi text.
- Cebuano had a single (ASCII) encoding, whereas Hindi text appears on the web in numerous encodings, forcing us to spend much of our time developing encoding converters to transliterate Hindi texts into a standard encoding.

Resource	**Cebuano**	**Hindi**
Text (words)	250K	>100M
Bilingual text (words)	130K	>5M
Lexicons (headwords)	25K	30K
Text annotated for named entities (words)	10K	430K
Text tagged for topic detection (documents)	None	>2200
Texts with summaries (documents)	None	25
Morphological parsers or stemmers	1	4

Table I. Major Resources for Cebuano and Hindi

We also summarize some important factors that affected our work in more or less the same way for Cebuano and Hindi:

- We did not find sufficient bilingual news text for our needs in either language; we were therefore forced to create translations using translation agencies or other means.

- While we found lists of entities in both languages, it was still necessary to do manual annotation of named entities in texts.
- Manual annotation was also required to provide training data for a range of technologies.

The fact that the two exercises targeted quite different languages for purposes of NLP gives us some confidence that we can base our future resource collection and creation efforts on these experiences.

We should however note that we did not face a major problem from morphology for either language; we had a morphological parser for Cebuano and we found one for Hindi (and one site developed a simple stemmer). Had we faced a language with a more complex morphology, and were unable to find an existing morphological parser, we would have had to expend considerably more effort in building a parser (or stemmer). While some effort has gone into machine-learning approaches to morphology [Maxwell 2002], the state of the art is not up to the automatic creation of morphological parsers or stemmers for languages with any degree of complexity in their morphology.

9. FUTURE: A CALL FOR COLLABORATION

Research on the rapid porting of linguistic technologies to new languages is crucial, as it helps determine the most efficient porting methods and encourages cost-benefit analyses of the types and sizes of linguistic resources necessary. However, it will always be preferable to avoid the scramble inherent in rapid porting by preparing and providing core linguistic resources in advance of need. Therefore we propose an initiative to begin collecting the resources necessary to develop critical language technologies in all target languages.

Necessary resources, varying both with the technology and the target language, are open to negotiation. However we propose that a core include significant bodies (minimally 100,000 words) of electronic text and parallel text, medium-sized translation lexicons (10,000 words), and entity databases and texts tagged for entities and topics. Such resources support information access technologies that work with text and are simple enough that it should be possible to locate them rapidly for a large number of languages. Although there are numerous other desirable resources, we propose that this initiative begin with attainable goals in order to maximize the probability of early success. The choice of target languages is similarly open to negotiation, but we propose that the work continue targeting larger languages in priority order.

Nevertheless, it is clear that LDC cannot tackle this task of language documentation for any large number of languages, even with the help of the other sites involved in the TIDES surprise language exercise. Accordingly, we invite a global participation in this effort. Participants would define their local priorities in collaboration with other interested groups working in the same or related languages. We are encouraged by recent efforts of European initiatives like ELSNET and ENABLER in this regard, and hope to work with these groups to develop approaches and resources that are both complementary and compatible.

Whatever the approach adopted for the documentation, collection, development, and distribution of resources for any particular language, we propose four principles to coordinate the effort:

(1) Individual participants will conduct language resource surveys and will identify, collect, and further develop linguistic resources, making the results available to the whole group. Access to the group survey results would be contingent upon substantive contribution to the effort.

(2) Although many of the targeted resources are already available on the Internet for research purposes, world-wide resource providers will need to engage data creators in intellectual property negotiations in order to secure distribution rights and then

distribute resources through existing channels. This will add value to the raw resources by creating corpora that are stable, consistently structured, and capable of being used for a variety of purposes.

(3) Participants are encouraged to use standard, freely available tools (such as the annotation tools we describe above) in order to encourage ongoing resource creation in a framework that promotes easy exploitation of its results by the widest possible audience.

(4) The resources created through this process should be made available to the world-wide community of researchers using an archival distribution method, and indexed so that other researchers can find the resources and make use of them, while respecting intellectual property rights [Bird and Simons 2003].

Our experience with the many-language resource survey, the rapid collection, development, and dissemination of linguistic resources and the highly collaborative framework of the surprise language exercise lead us to believe that a broader, more ambitious, effort is not only possible but obligatory, given the current state of language technologies and the focus of technology programs world-wide.

REFERENCES

AL-ONAIZAN, Y., CURIN, J., JAHR, M., KNIGHT, K., LAFFERTY, J., MELAMED, D., OCH, F-J., PURDY, D., SMITH, N. A., AND YARMOWSKI, D. 1999. Statistical machine translation. Final Report. Johns Hopkins University Text, Speech, and Dialog Workshop 1999. http://www.clsp.jhu.edu/ws99/final/Stat_Machine_Translation.pdf.

BEESLEY, K. R. AND KARTUNNEN, L. 2003. *Finite State Morphology*. Stanford, CSLI.

BIRD, S. AND LIBERMAN, M. 2001. A formal framework for linguistic annotation. http://agtk.sourceforge.net.

BIRD, S. AND SIMONS, G. 2003. Seven dimensions of portability for language documentation and description. *Language 79* (2001), 557-582.

BUNYE, M. V. R. AND YAP, E. P. 1971. *Cebuano Grammar Notes*. University of Hawaii Press, Honolulu.

BYRNE, W., HAJIC, J., IRCING, P., JELINEK, F., KHUDANPUR, S., MCDONOUGH, S., PETEREK, N., AND PSUTKA, J. 1999. Large vocabulary speech recognition for read and broadcast Czech. In *Proceedings of the Text, Speech, and Dialog Workshop* (1999).

CHINCOR, N. 1997. MUC-7 named entity task definition version 3.5. http://www.itl.nist.gov/iad/894.02/related_projects/muc/proceedings/ne_task.html.

CIERI, C. AND LIBERMAN, M. 2002. TIDES language resources: A resource map for translingual information access. In *Proceedings of the Third International Language Resources and Evaluation Conference* (Las Palmas, Spain, May-June 2002).

ELSNET. 2003. European Network of Excellence in Human Language Technologies website. http://www.elsnet.org

ENABLER Network. 2003. European National Activities for Basic Language Resources Network website. http://www.enabler-network.org.

FURUI, S. 2001. From read speech recognition to spontaneous speech understanding. In *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium* (Nov. 27-30, 2001, National Center of Sciences, Tokyo), 19-25. http://www.afnlp.org/nlprs200/pdf/inv-03-01.pdf.

GHANI, R., JONES, R., AND MLADENIC, D. 2001. Mining the web to create minority language corpora. Presented at the *Proceedings of the 10th International Conference on Information and Knowledge Management (CIKM)*. http://citeseer.nj.nec.com/ghani01mining.html.

GRIMES, B. (ed.) 2003. *Ethnologue: Languages of the World.* 14th ed. SIL International, Dallas, TX. http://www.ethnologue.com.

MAXWELL, M. (ed.) 2002. *Workshop in Morphological and Phonological Learning: Proceedings of the Workshop.* Association for Computational Linguistics, New Brunswick, NJ. http://cogsci.ed.ac.uk/sigphon/CPapersSP.html.

KIRCHOFF, K. et al. 2002. Novel speech recognition models for Arabic, Final presentation of the Johns Hopkins University Text, Speech, and Dialog Workshop (Aug. 21, 2002). http://www.clsp.jhu.edu/ws2002/groups/arabic/asr-final.ppt.

Linguistic Data Consortium. 2003. TIDES project web site. http://www.ldc.upenn.edu/Projects/TIDES.

Linguistic Data Consortium. 2003. Surprise language project web site. http://www.ldc.upenn.edu/Projects/SurpriseLanguage.

PSUTKA, J., IRCING, P., PSUTKA, J. V., RADOVIC, V., BYRNE, W., HAJIC, J., MIROVSKY, J., AND GUSTMAN, S. 2003. Large vocabulary ASR for spontaneous Czech in the Malach project. Manuscript submitted to the *Proceedings of the European Conference on Speech Communication and Technology* (EUROSPEECH).

SIL International. 2003. *Ethnologue: Languages of the World*. http://www.ethnologue.com/.

STRASSEL, S. 2003. Surprise language annotation specifications. http://www.ldc.upenn.edu/Projects/SurpriseLanguage/Annotation.

WAYNE, C. 2002. Human language technology: TIDES, EARS, Babylon. In *Proceeding of the 2002 DARPA Tech Symposium* (Anaheim, CA, July 30-Aug. 2, 2002). http://www.darpa.mil/DARPATech2002/presentations/iao_pdf/speeches/WAYNE.pdf.

RECEIVED AUGUST 2003; REVISED OCTOBER 2003; ACCEPTED OCTOBER 2003