MULTILINGUAL TEXT RESOURCES AT THE LINGUISTIC DATA CONSORTIUM

David Graff, Programmer Analyst
and
Rebecca Finch, Research Coordinator

Linguistic Data Consortium
University of Pennsylvania
441 Williams Hall
Philadelphia, PA 19104-6305

ABSTRACT

The Linguistic Data Consortium (LDC) is currently involved in a major effort to expand its multilingual text resources, in particular for machine translation, message understanding and information retrieval research. The main sources for data acquisition are governmental and international organizations, newswire services, and diverse publishers. This paper describes some of the research that is being done to identify potential resources, discusses some of the process involved in negotiating the broadest possible access to the material for the human language technology research community, and identifies key issues and considerations in transducing the text into common and well documented formats.

1.0 GOVERNMENTAL AND INTERNATIONAL ORGANIZATIONS

The LDC has acquired archives from the United Nations, the Hansard corpus of transcriptions from Canadian parliamentary debates, hearing and proceedings, and is working to acquire multilingual archives from three subsidiaries of the United States Information Agency, Voice of America, Radio Marti, and Radio Free Europe/Radio Liberty. A Defense Department publication, *Military Review,* has also committed a four year archive of English, Spanish and Portuguese text.

1.1 The United Nations Multilingual Corpus

The LDC has acquired a five-year electronic text archive of English, French and Spanish documents of public record from the United Nations. The documents include the proceedings, resolutions and reports of the General Assembly, the Security Council, UNICEF, the Economic and Social Council, and numerous othe committees, commisions and councils within the UN. The UN has also agreed to provide its Russian, Arabic and Chinese archives for the same period. Russian and Arabic text have been received from the United Nations in Geneva; the LDC has yet to receive Russian and Arabic from the UN in New York, and no Chinese has been delivered to date.

About 60% of the 2.5 gigabytes of English, French and Spanish data that was delivered to the LDC was identified as being parallel in all three languages; the remaining 40% is made up of material that exists in only one or two of the three languages. The LDC has not yet begun to determine the extent to which the Russian and Arabic archives are parallel with the material already received and processed.

The UN texts were created and archived on Wang VS computer systems, using an obsolete Wang word processing program. The tapes delivered to the LDC were copied from 80 megabyte removable disk packs by means of Wang BACKUP. Each of the Wang programs used its own file formatting scheme, which had to be reverse-engineered at the LDC so that programs could be written to extract the actual text data from the tapes. The LDC's efforts to decipher the WP character encoding, format control codes and file structure were helped substantially by Dominique Petitpierre of ISSCO, as well as by the technical support at Wang Office Systems.

The process of sorting the text files into parallel sets and assigning index numbers that identified each document in each of its translations consisted of the following steps.

1) Complete document lists were compiled for each year, containing all entries from all languages for that year.

2) Each list was sorted with respect to the UN "job-number" associated with every document in the list. The UN assigns job-numbers to documents in roughly chronological order. Each translation of a document is supposed to have a unique job-number, although this convention breaks down in a significant number of cases. It is generally the case that the various language versions receive their numbers in a fixed sequence, with English being first, French second and Spanish fourth. As a result, within a given year a 3-way

parallel set for a given document should have job numbers of the form yy-<x> (English), yy-(<x+1) French, and yy-(x+3) Spanish.

There were numerous files in the archives that had received no numbers; there were some cases where the numeric sequence had nothing to do with parallel relations across languages, and there were even some cases where very different files were assigned the same number. For this reason, we have derived only a 60% yield of parallel data from the archives at this point.

3) Each list was passed through a procedure that assigned a sequence number to every English entry in the list; the sequence number started at 00001 for each year, and was simply incremented for each English file. A corresponding numeric field of 00001 was applied to every French and Spanish entry in the list.

4) A specialized string-matching process was applied to each list to pull out the clear cases of parallel entries: for each English entry in the list, its sequence number was copied into the corresponding field of a subsequent French or Spanish entry if such an entry was found to be parallel to the English. Parallel entries were moved to a list of "matches," English entries for which no parallels were found, and other entries that did not match an adjacent English entry, were moved to a residual list.

5) Each residual list was passed through an interactive process in which a user inspected likely candidates for parallel sets that had failed the previous string-matching method. The user had the option of inspecting the actual contents of the text files listed in the candidate sets. If, after inspecting the list entries (and possibly the files themselves), the user determined that two or three of the candidate entries constituted a parallel set, these entries were selected for addition to the lists of matches. The remainder was again left in a residual file.

6) The lists of matches were used to drive a process that copied each parallel file into its appropriate place in the directory tree shown above, according to its language, year, and sequence number.

Note that because some English documents did not have any parallels in French or Spanish, there are gaps in the sequence numbers within each year. In other words, only those English files having parallels have been included in this release. Additional parallels are likely to be found as the source data are examined more closely, and these will be distributed later to fill in some portions of gaps in the numeric sequence.

To aid users of this corpus in finding the parallels for each English file in this release, a table has been provided for each year for each of the three language directories. The tables list every English file present in the release by its sequence number, and indicate whether the French or Spanish (or both) versions are also present for that file. For the years 1988-1993, 22,110 English documents, 20,350 French documents, and 14,773 Spanish documents are included.

The English, French and Spanish texts were transliterated to the ISO 8859-1 (Latin1) character set, an 8-bit encoding system in which accented characters of European languages (and some specialized symbols) are provided in the upper half of the 256-character table. Common 7-bit ASCII, or ISO 646, occupies the lower half of the table.

In addition, the various WP text formatting control codes (such as line-centering, underlining, indentation, tab-stop settings, etc.) were preserved in the form of Standard General Markup Language (SGML) tags. Considerable care was taken to ensure that the resulting text files are fully SGML compatible and parsable. Important assistance in this effort was received from David Mckelvie of the HCRC in Edinburgh, Scotlant, for providing a critique and verification of some extracted samples, and for creating a complete SGML Document Type Definition (DTD) and character set specification, which will be distributed with the data when it is published.

It is not clear at present what solution will be adopted for character encloding in the other three languages. The LDC is currently looking for widely used methods for encoding Arabic, Russian and Chinese that also contain roman characters as a proper subset. If these are available, SGML markup on these three sets of files can probably be accomplished in a fairly normal way. If not, then we will have to look at ways for revising SGML for use in each context, or we will have to use a different method of markup. We welcome suggestions from users of this type of material as to ways of moving ahead with this.

As a final task, the LDC returned two text archives to the United Nations, one to Geneva and one to New York, composed of the entire set of documents in three languages organized in a manner directly parallel to the organization of the original WANG archives, but converted to WordPerfect format.This was a service the LDC performed for the UN in exchange for their contribution of the archives for research purposes. The Geneva archive was converted from WANG WP to WordPerfect by Dominique Petitpierre at ISSCO in Geneva, using a utility program called Aladdin Transfer. The New York archive was converted to WordPerfect by running WordPerfect's Intellitag software against the files once they had been tagged with SGML.

Once the processing of the last three languages is complete, the LDC intends to return to the United Nations to negotiate for access to other data resources. These include the archives from the UN Publications Department and UNESCO archives in Paris.

1.2 The Hansard Corpus

Two large bodies of data composed of French and English transcripts from proceedings of the Canadian Parliament have been donated to the LDC by IBM and Bellcore. The IBM part of the corpus totals 46.3 million words of French and 38.6 words of English, all of which are sentence aligned; the sentences have been systematically scrambled to prevent the corpus from being used as an information resource. The Bellcore Hansard consists of roughly 60 million words;

since the file naming conventions used do not permit distinguishing of the French from the English in a large portion of the corpus, exact counts of French and English words have not yet been done.

The LDC will publish the Bellcore part of the corpus sometime in the spring of 1994.

1.3 United States Information Agency

The LDC is working to gain access to multilingual material from three subsidiaries of the United States Information Agency (USIA), Voice of America (VOA), Radio Marti (RM) and Radio Free Europe/Radio Liberty (RFE/RL).

VOA is an information organization with international headquarters in Washington, D.C. The broadcasting is done by 46 language services, which are semi-autonomous radio station staffs organized by linguistic coherence or regional orientation. Decisions about where VOA should broadcast to are made by the National Security Council. Total broadcasting time is 1000 hours per week in 47 languages. Each staff adapts a centrally produced English language product through translation, paraphrase or original writing for dissemination to its particular area. VOA broadcasts to other countries about the United States. The languages it broadcasts in are Albanian, Amharic, Arabic, Armenian, Azerbaijani, Bangla, Bulgarian, Burmese, Cantonese, Creole, Croatian, Czech, Dari, English, Estonian, Farsi, French, Georgian, Hausa, Hindi, Hungarian, Indonesian, Khmer, Korean, Kurdish, Lao, Latvian, Lithuanian, Mandarin, Pashto, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Slovene, Somali, Spanish, Swahili, Thai, Tibetan, Turkish, Ukrainian, Urdu, Uzbek and Vietnamese. All VOA broadcasts are analog recorded; the LDC has not yet been able to ascertain how far back the recordings go. The VOA has a computer readable English language archive of scripts that is about two gigabytes; it dates back to October 1990. The size of the archives being compiled by the individual language services is unascertainable at this time, because no uniform archiving policy is currently in effect.

Radio/TV Marti is a separate organization that broadcasts to Cuba 24 hours per day in Spanish only. Much of its product is extemporaneously produced, that is, no scripts are written prior to broadcasts. Spokespeople for the service state that they have a large electronic text archive of Spanish only that consists of material transcribed for use as an information resources; it includes, for example, all of the speeches that Fidel Castro has ever given. Radio Marti's broadcasts are digitally recorded.

Radio Free Europe/Radio Liberty is a USIA subsidiary with international headquarters in Munich, Germany. It broadcasts in 29 languages; its major regional focus is Eastern Europe. It broadcasts to other countries about subject matter other than the United States (i.e., it broadcasts to Russia about Russia).

There are two obstacles that must be overcome in order for the USIA's resources to be made available to the research community. One is a political obstacle, the other a technical one; both are described here. To gain the right to access the materials created by VOA and the Cuban Broadcasting Corporation, the LDC and its host institution, the University of Pennsylvania, must succeed in getting a congressional amendment through Congress that will exempt it from a congressionally mandated ban on dissemination of these USIA materials in the United States. The ban was formulated in the Smith-Mundt Act that created the USIA in 1948, in a period of time when many congressional representatives and the public still remembered the power and influence of state run propoganda organizations built in countries like Nazi Germany and the Soviet Union. Fearful of what might happen should an agency like the USIA be allowed to operate within the United States without restriction, Congress forbade distribution of any of its materials within the United States. A procedure was put into place, however, that allowed for the release of individual movies, videos, photographs or scripts; it requires that Congress pass an amendment authorizing the release of such materials. In the forty-four years since Smkith-Mundt was passed, thirty-two exceptions to the dissemination ban have been allowed.

None of the previously passed exceptions is as all encompassing as the exception that the LDC is requesting, however. All of the previous exceptions were granted for one movie, or for all of the photographs taken by one USIA employee, or for all USIA films, scripts and photographs that have President Kennedy in them. The LDC is requesting that it be given access to all of the electronically readable text in all of the 47 languages VOA publishes in, as well as its entire archive of recorded speech (assuming that the recordings are of high enough quality to be usable). It is also requesting the same access to Radio/TV Marti's archives.

Congressional staff people from Harris Wofford's office have already made one attempt to get an amendment attached to the USIA authorization legislation. This failed, primarily, we believe, because we did not have sufficient time to gather support from LDC's members and the government organizations that LDC distributes data for (ARPA, NIST and the NSF). At this point, Wofford's staff is preparing to introduce the legislation as an amendment to the Foreign Aid authorizations bill, which will be introduced into committee in April or May of this year. The LDC will be asking its members to contact their appropriate Senate and congressional staff people to explain how important these resources are from the standpoint of the research community.

Once the political obstacle is overcome, the technical one will have to be addressed. Scripts for all VOA broadcasts except those in Tibetan and Cantonese are produced with Xerox's Globalview Workstation text processing product, in which character encoding is based on a proprietary 24-bit scheme that is not Unicode compatible. The mapping tables are not readily accessible from Xerox (it is not even clear if they exist at all), and documentation for GlobalView's file formats is basically a dump of the runtime heap, so that in essence an external program to unscramble it will be equivalent to GlobalView itself. In order to make the multilingual text resources available, some means has to be found to convert the text created with GlobalView into Unicode with some well-defined and easily parsed markup.

Such conversion software will have to be developed to run in batch mode if the entire archive is to be made accessible.

1.4 *Military Review*

Military Review is a publication of the U.S. Department of of the Army produced in Fort Leavenworth, Kansas. The journal is published in English eleven times per year, and six times per year in Spanish and Portuguese. The Spanish and Portuguese editions are primarily translations of articles from the English version. The editor of the Latin American edition has authorized transfer of 2250 megabytes of data to the LDC; however, since the files are in Interleaf format, a large percentage of this will be graphical data, so ultimate text yields cannot be estimated at this time.

2.0 FOREIGN LANGUAGE NEWSWIRE

One of the most important sources for large quantities of foreign language text being developed by the LDC is news services, both domestic and foreign. Currently, the LDC is receiving multilingual text from both Associated Press and Reuters.

2.1 Associated Press

Associated Press (AP) Worldstream is an output that is an amalgamation of all the AP-produced international services. The English language copy usually originates in or is of interest to areas outside the United States; the services also produces copy in French, German, Swedish, Dutch and Spanish. Worldstream is sent in a wire service transmission envelope developed by the American newspaper Publishers Association (ANPA) for 1200 baud transmisions. The character set is not the ANPA standard character set, however, but rather a modified version of the IBM code page 850 character set that allows representation of all the accented characters of the languages contained in the service. Because the output is produced by an editorial staff that is centrally located in New York City, the expectation is that a large percentage of the material will be translated and therefore parallel. The LDC has not yet determined to what extent this is the case, however.

2.2 Reuters

Reuters Spanish Language Service originates in Argentina; Reuters Latin American Business Report, which is advertised as being composed of Spanish, Portuguese and English, originates in Brazil. These two services are complemented with Reuters World Service, which provides English language text, much of which is paraphrased or translated from or to the other languages. The LDC receives the data via a satellite dish installed at the University of Pennsylvania. As with the AP data, Reuters' services are sent in an ANPA transmission envelope, but three modified ASCII character tables are used to transmit basic newspaper text, graphics characters, and complex numerical data. Again, the LDC has only recently started trying to determine to what extent these services will actually provide parallel resources.

2.3 Agence France Presse

Beginning sometime in the first quarter of 1994, Agence France Presse (AFP) and Xinhua News Agency will begin transmission of their news service products to the LDC. The AFP output will be received via satellite, and will be composed of text in six languages: English, French, German, Spanish, Portuguese and Arabic. The English, French, German and Spanish services use a unique 7-bit ASCII character set developed by AFP. The Portuguese service uses a variation of the IBM code page 850 character set. The Arabic service uses a proprietary character set developed by AFP for DOS.

The English and French services are sent in the ANPA transmission envelope; the German and Portuguese services are sent in the transmission enveloped designed by the International Press Telecommunications Council (IPTC). This is a variation of the ANPA guidelines that takes into account the technical and linguistic differences between countries and is designed for use in numerous languages and services.

The degree to which materials in the separate language newswires can be expected to be parallel is not known at this point.

2.4 Xinhua News Agency

The Xinhua News Agency is unlike most of the other news agencies in several important respects. First, it is a government news service whose product is centrally produced in Mandarin; it produces text in five other languages that is entirely translated (not paraphrased or adapted, as is the case with the other news service agencies we have been discussing). The LDC will receive Mandarin in both GB and BIG 5 format, Spanish and English at the start of its contract with Xinhua. The service also translates all of its articles into French, German and Arabic, but these services are not transmitted to the U.S. on a regular basis. The LDC is negotiating with Xinhua to receive these latter three languages on tape or diskette.

2.5 Other News Services

A number of other newswire sources are under negotiation or consideration. **YONHAP (Korean Press Agency)** is a Korean language service that provides about 130 articles per day in Korean and 10-30 articles in English. This is one instance where the vendor has clearly informed us that the English bears little, if any, relation to the Korean material, so the LDC is negotiating for only the Korean text, which uses only Hangul characters (no Mandarin). The data is transmitted in 5-bit ASCII in a nonstandard (non-ANPA) transmission envelope.

dpa (German News Agency) produces international news and feature services in German, English, Spanish and Arabic. The LDC is negotiating to receive German and English, although where the service will be received is in question at this point. The English language service is available via Associated Press' Datafeature network, and requires only the installation of a small satellite dish and a small monthly communication charge. The German service is available in the United States but only in New York or Washington. Delivery of that service to the LDC at the University of Pennsylvania will require the installation of a dedicated telephone circuit from either city. The cost for installing such a line is $3,500+, and the monthly communication charge is almost $800. Clearly, this makes the expense of this service unreasonable. The LDC is therefore, looking for a logical collection point in Europe. dpa's service is transmitted in IPTC format.

Kyodo News Service provides a financial news service in Japanese that is primarily oriented toward Japanese banks and corporations in the United States.

ITAR-TASS and Interfax are Russian news services headquartered in Moscow. ITAR-TASS was the official Soviet government news agency until recently. Interfax was formed in 1989, during the period when Glasnost and Perestroika were developing, and advertises itself as a news service that represents the ideas of that period. Both of these services provide substantial material that is written in Russian and translated to English. ITAR-TASS materials are transmitted to the United States via the AP Datafeatures network satellite; the data is received on a PDP-8 computer and then transmitted to ITAR-TASS's client base. The service also produces transcripts of debates and proceedings in the Russian parliament that are available in the transcribed Russian and translated English.

Interfax publishes fifteen daily and weekly reports in Russian and English using Microsoft Word on IBM PCs. In addition to ten daily and weekly business oriented publications, Interfax also has a 400 megabyte archive of Russian and English text available.

In addition to the above news services, the LDC is investigating fifty international news services that offer text in languages that are difficult to acquire; all of these will require finding sites in Europe, Hong Kong, and perhaps other places in order for the data to be received.

3.0 EUROPEAN CORPUS INITIATIVE

The ECI disk will be published and distributed by the LDC sometime in the first half of 1994. The disk will contain somewhere between 41-48 subcorpora (the final number depending on permissions) in 26 languages with about 92 million (lexical) words. Nine of the subcorpora are multilingual with parallel data, including the Corpus of International Labour Organization reports.

4.0 SPANISH LANGUAGE JOURNALISTIC TEXT

A broad range of foreign language publications originating in North, Central and South America is being contacted for text acquisitions. El Norte, the Pulitzer Prize winning daily newspaper from Monterrey, Nueva Leon in Mexico, is making 200MB of text available. Excelsior, a daily newspaper from Mexico City, and Nueva Provincia, a daily from Bahia Blanca, Argentina have agreed to make unspecified amounts of text available. Contacts with Mexican, Central and South American daily newspapers will continue to be made for some time.

The College Division of McGraw-Hill, Inc. is contributing text from a number of Spanish language readers and textbooks.

5.0 ENGLISH LANGUAGE RESOURCES

In addition to the focus on multilingual resources, the LDC is investing considerable time and money in English language text resources as well.

The LDC has filed a Freedom of Information Act petition with the United States Justice Department for acquisition of the public domain portions of the JURIS database, which total about 7.5 gigabytes of data. The Justice Department has stated that we will receive the data once it has permissions from the 29 agencies and 9 Justice departments to distribute the data.

The LDC is also expanding its newspaper holdings. The Dow Jones Information News Service (a financial news service), the *Wall Street Journal*, and the New York Times News Service (which is composed of 300,000 words per day of the *NY Times* newspaper and text from several North American news services) are all in the process of contract negotiations with the LDC. The Los Angeles Times and Washington Post News Service is considering the LDC's request for access to its data for research purposes.

The LDC has also acquired permissions for the *Philadelphia Inquirer* newspaper from Knight Ridder, which also owns the *San Jose Mercury News,* which was used in the Tipster program last year. We are also requesting permissions from two other Knight Ridder papers, the *Detroit Free Press* and the *Miami Herald.*