TIDES Language Resources: A Resource Map for Translingual Information Access

Christopher Cieri, Mark Liberman

University of Pennsylvania, Linguistic Data Consortium
3615 Market Street, Philadelphia, PA 19104-2608 U.S.A.
{ccieri, myl}@ldc.upenn.edu

Abstract

Continuing improvements in human language algorithms, coupled with improvements in digital storage and processing, inspire growing confidence in multilingual information access systems. Systems exist to transcribe broadcast news, segment broadcasts into individual stories and sort them by topic. These technologies, useful in isolation, are now being combined to produce intelligent multilingual systems. DARPA TIDES combines technologies in detection, extraction, summarization and translation to create systems capable of searching a wide range of streaming multilingual text and speech sources, in real time, to provide effective access for English-speaking users. The broad scope of tasks and languages in programs like TIDES demands close coordination of research and shared resources. These resources includes large collections of raw text and speech; translations and summaries; annotations of topics, named entities and relations, syntactic structures and propositional content; lexicons; annotation specifications and protocols; and distribution formats and standards. The TIDES program has initiated ambitious attacks on difficult problems, with linguistic resources matched to the needs of each piece of the overall research enterprise. This paper will describe the coordinated language resources being created under the TIDES aegis.

1. Introduction

The past 15 years of research in Human Language Technology (HLT) have shown that effective linguistic technology is based on statistical modeling of large amounts of linguistic data, and that the most reliable way to improve linguistic technology is to improve the linguistic resources upon which it is based. For familiar languages, improved HLT technology depends on order-of-magnitude increases in the underlying text and speech corpora, while porting HLT technology to new languages requires creation of similar-sized linguistic resources for them. As database sizes increase, new research methods come into play; at the same time, in order for database sizes to increase, new and more efficient methods of collection and creation, based on earlier research results, are needed. Today's key HLT research challenge is to create and digest linguistic resources on a significantly larger scale than ever before.

The DARPA program in Translingual Information Detection Extraction and Summarization (TIDES) aims to enable users to find and interpret needed information efficiently regardless of language or medium. TIDES research areas include information detection, extraction, summarization and translation; researchers in the program work on one or more areas, their intersection or integration into end-to-end systems. TIDES core languages are English, Mandarin, Arabic; second tier languages are Korean, Spanish and Japanese. The primary medium is text though this includes speech recognition output.

2. Data in Human Language Technology

Recent improvements in HLT algorithms, coupled with on-going cost/performance improvements in digital storage and processing, inspire growing confidence in the future value of HLT technologies. Systems exist to transcribe broadcast news, segmenting programs into stories and sorting them by topic. There is ongoing research into information extraction from text, question answering, document summarization and machine translation. These technologies, useful in isolation, address current needs even more effectively when combined. Several DARPA-sponsored common-task research programs now focus on extending HLT to new languages, and combining technologies to produce multilingual intelligent systems. The DARPA TIDES program combines technologies in detection, extraction, summarization and translation to create systems capable of searching a wide range of streaming multilingual text and speech sources, in real time, to provide effective access for English-speaking users.

Extension of HLT technology to wider coverage of the world's languages raises some new research problems, which in turn require new kinds of linguistic data. For example, treatment of collections of related languages and dialects requires new kinds of adaptive algorithms, whose development must be based on new kinds of multi-dialect resources.

The Linguistic Data Consortium (LDC) was founded in 1992 at the University of Pennsylvania, with seed money from DARPA, specifically to address the need for shared language resources. Since then, the LDC has created and published more than 209 linguistic databases, and has accumulated considerable experience and skill in managing large-scale, multilingual data collection and annotation projects. Responding to the need for more data in a wider variety of languages with more sophisticated annotation, the LDC has established itself as a center for research into standards and best practices in linguistic resource development, while participating actively in on-going HLT research.

In the context of DARPA TIDES we have begun to develop methods for creating and using linguistic resources on a larger scale than the HLT research community has previously undertaken, both in terms of the amount of material per language, and the number of languages covered. This work requires close collaboration with other DARPA HLT researchers.

3. The DARPA TIDES Program

The broad scope of tasks and languages in programs like TIDES demands close coordination of research and shared resources. These resources includes large collections of raw text and speech; translations and summaries; annotations of topics, named entities and relations, syntactic structures and propositional content; lexicons; annotation specifications and protocols; and distribution formats and standards. TIDES has initiated ambitious attacks on difficult problems, with linguistic resources matched to the needs of each piece of the overall research enterprise. LDC's role in TIDES is to coordinate resources for the multiple common-task, metrics-based technology development and evaluation projects.

A recent survey of TIDES participants showed that the most often used resources were the annotated corpora, such as those created for TDT and TREC, followed closely by bilingual lexicons, Treebanks and parallel texts. Participants saw the greatest future need for more and bigger parallel corpora, more and better bilingual lexicons, and more and bigger treebanks and proposition banks.

This survey concluded that resource development should focus on TIDES core languages and give priority to:

1) direct support for common tasks

2) data essential to the task definitions in detection, extraction, translation, summarization, for example continuations of the TREC, TDT and ACE corpora

3) indirect support for common tasks by providing data needed by most algorithms for example bilingual dictionaries for cross-language information retrieval

4) background resources used to develop better system components for example texts, bitexts, treebanks, proposition banks

TIDES resources can be categorized roughly as follows.

4. Gigaword News Text Corpora

Very large scale text databases of roughly one billion words per language supports robust statistical modeling, and provides raw data to be annotated as evaluation resources for all research groups. In addition, comparable corpora of this size permit statistical MT research even in the absence of parallel (translated) text. Current target languages are English, Mandarin and Arabic. The worldwide web contains text in such volume in many languages. However, web data is subject to copyright that constrains the ability to share such material across research groups. Furthermore the dynamic nature of web pages makes them undesirable as evaluation resource in their natural state. One solution is for a central archiving agent to harvest web material and to establish a legal basis for broad and ongoing research access to the resulting archive. News services are another source of large volumes of text data, and are accustomed to licensing it for external use. Here again, however, a central archiving agent is essential to establish legal foundations and to apply common standards. In each language, we pursue an appropriate strategy to guarantee adequate text access for the research community, in a timely and cost-effective fashion.

There has been considerable progress in building the Gigaword News Text corpora. LDC has acquired rights to distribute more than 1.4 billion words of English, 1.5 billion characters of Chinese and nearly 500 million words of Arabic. Work is underway to format these corpora for broad distribution.

5. Broadcast News

Broadcast news is an important domain for human language technology both because of the inherent interest in systems that can process news and because its vocabulary is expansive and touches most areas of everyday life. LDC has collected and annotated broadcast news since the mid-1990's for DARPA's common task research projects.

TIDES needs broadcast news in much greater volume, from more sources and over a longer duration than any previous program. This initially involves broadcast radio and television sources and Internet based distributions (webcasts) in English, Chinese and Arabic with possible additions in the coming years.

The problem of finding legal/technological models for creation and distribution of research archives of BN has become an acute one. Standard license fees for broadcast materials, even for non-profit or educational use, are established at between $10 and $80 per second. In the past, LDC has managed to get licenses for thousands of hours of BN at typical costs of $100 per hour. This worked well when research needs were for tens or hundreds of hours. Now, although it is two or three orders of magnitude below the "list price", it is still too much to pay as the research needs escalate to tens of thousands of hours. Even worse, there are serious delays and even exclusions entailed by the arduous process of IPR negotiations in this domain, especially for broadcasts from other countries.

Whatever the legal model, LDC will research best practices for BN collection, including identifying sources with appropriate coverage and availability, selecting those that can legally be recorded and archived for research purposes, developing mechanisms to capture the data efficiently and reliably, implementing a local version of the capture systems and testing it on long-term collections. We will carry out these collections completely in the digital domain throughout the duration of the project, yielding broadcast news collection of up to five years in time depth. Both the description of the collection mechanism and the resulting data will be made available generally.

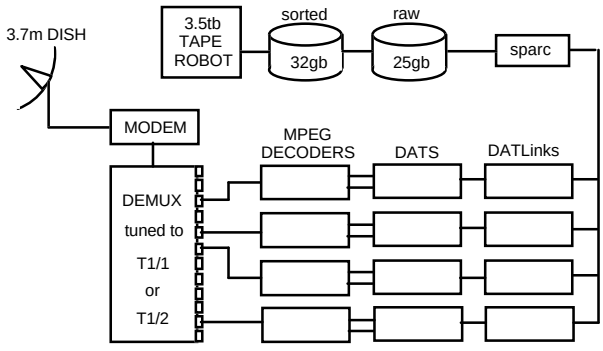

Figure 1: Current configuration of multi-channel broadcast news capture facility

The diagram above shows our system for collection of Voice of America and other broadcasts, which is typical of the problem of multi-channel digital data capture. Our satellite antenna captures the digital, multi-channel, broadcast signal directly from Voice of America's worldwide distribution network. That signal is de-multiplexed into up to 24 simultaneous channels of MPEG audio and video. A series of digital audio tape recorders capture the output of the MPEG decoders sampling the audio at rates up to 48KHz. The digital audio passes through the Townshend DAT Links and onto spinning disk managed by a small Sun workstation. Raw audio is written to disk in daily clusters where staffers review it for signal quality and sort it by language. The sorted data is written to near-line storage maintained on a 3.5TB tape robot. Broadcast news capture from other sources – cable TV, consumer satellite and the broadcast airwaves will require similar programmable data capture systems that we propose to plan, deploy and evaluate over the course of this project, in collaboration and consultation with other sites that do similar sorts of data capture. Work on the all-digital broadcast news capture facility is still in the planning stage.

6. Parallel Text

Parallel text describes documents that have been translated from the original source language into one or more target languages. Statistical machine translation technology is based on relatively large volumes of parallel text – at least tens of millions of words for good coverage. Available algorithms require parallel texts to be accurately aligned at the sentence level. Parallel text of adequate quantity and quality is hard to find, and as a result significant experiments have only been attempted in a handful of languages. LDC has experience harvesting parallel text from the WWW, acquiring parallel-text archives from a variety of sources, and managing intellectual property rights for the distribution of these archives to researchers. To solve the problem of locating parallel text, LDC has created technology called Bilingual Internet Text Search (BITS). Recently BITS has unearthed large repositories of parallel Chinese-English, German-English and Korean-English text on the worldwide web.

BITS was developed to find, collect and align parallel text from Internet. Other researchers, especially Resnik and others at the University of Maryland, have proposed indexing the web to create a list of URLs pointing to translation pairs. However, URLs are not durable references, and so this solution does not create databases that can be used over time, as is required by a typical DARPA common-task research paradigm. Stable access is required in order for research progress to be documented, by comparing the performance of different algorithms on the same data, both across research sites and across time. However, given a durable Internet archive, accessible to researchers in the way that library archives are – such as is promised by http://archive.org -- Resnik's approach may become a practical one. LDC is cooperating with Resnik and others to explore all feasible approaches to finding and obtaining access to parallel text for TIDES research.

This includes improving the performance of web spiders specialized for finding parallel text, as well as developing a range of legal/technological models for research access and/or distribution. Progress on harvesting Chinese-English parallel text from the Internet has been good. LDC has published the Hong Kong News, Hong Kong Laws and Hong Kong Hansards that were identified during a BITS search. The document-level alignment component of BITS has also been used to identify more than 19,000 translation pairs in the Chinese and English services provided by Xinhua news agency. The parallel Xinhua data will be distributed to TIDES research sites in 2002.

7. Bilingual Dictionaries and Morphological Analyzers

Multilingual lexical resources are critical for translingual technologies, not only translation but detection, extraction and summarization as well. A minimal lexicon for a new language relates word forms to a set of English glosses for each. Syntactic information, frequency information, pronunciation and so on are also useful. Often the set of word forms is too large to list conveniently, and it is more efficient to provide a list of word stems along with the set of affixes that may be added to each. Traditional methods of creating such resources are both expensive and time-consuming, involving human lexicographers working with concordances of text material and existing dictionaries. There are several ways to acquire lexical resources more efficiently. The first is simply to license such resources where already available. The second method is to apply machine-learning techniques to supplement human lexicographers, or to increase their efficiency. Both techniques have been applied in the HLT research community. However, providing lexicographic resources for new levels of performance and for larger numbers of languages will require improved techniques.

For many languages, a key lexicographic roadblock is the need for a system to analyze and synthesize word forms: a computational morphology. There are some promising ideas for the application of machine learning in this area. LDC has begun to confront this problem using an approach that has worked well in other HLT domains: to define and implement realistic test problems where algorithm performance can be evaluated quantitatively, and alternative approaches can be compared.

Our growing test bed will contain language data from approximately ten languages, chosen for their diverse morphological typologies. The data will include morphologically analyzed texts, dictionaries, and a computationally interpretable morphological grammar (useable by a parser and a generator, also provided). An interface will allow the use of this data in a variety of ways. For example, the texts can be treated as unannotated monolingual text or as glossed interlinear text; and the grammar can be used as a parser, or to simulate a human producing forms from a paradigm on demand.

By providing a standard set of morphological data on a variety of language types, it will be possible to fairly evaluate the capabilities and limitations of morphology learning systems which have been or will be developed, whether at our site or by other HLT researchers. Work on machine learning is progressing apace. Mike Maxwell's report appears in this same volume of proceedings.

8. X-Banks

Here we use the term X-Bank to refer collectively to Treebanks, Proposition Banks and related resources. Treebanks are collections of text that have been annotated to show the morpho-syntactic properties of sentences and their constituents. Treebanks, like those created for English, Chinese and Korean, are critical for development of many types of HLT technology. LDC is coordinating the extension and distribution of the extended Chinese Treebank and is creating a new Treebank in Arabic targeting one million words of source text. Each Treebank contains some parallel text to support research into transfer grammar.

Research efforts will continue to define the aspects of text meaning whose annotation is most effective in developing new technology, and also to steadily improve the efficiency of X-banking annotation by developing and applying new automatic techniques. The Arabic annotation team has made good progress on the part-of-speech tagging prerequisite to creating the Arabic Treebank. 130,000 words of news text has been part-of-speech tagged and hand-checked by human annotators. Of those, approximately 10,000 words have been syntactically annotated as of the time of writing.

9. Evaluation Resources, Detection

Research in information detection is often cast in terms either of identifying all documents that discuss a topic of interest or of clustering all documents according to their topic. Annotations that identify news stories or other documents types and consistently categorize them according to topic serve both kinds of research. LDC has created databases to support detection research in English, Chinese and Arabic under both the Topic Detection and Tracking program and the Text Retrieval Conference's Cross-Language Information Retrieval track (TREC CLIR).

TDT-4, currently under development is the most recent of the corpora created to support research in topic detection and tracking. TDT-4 is based upon data collected daily from October 2000 through July 2001 including eight English, seven Chinese sources and three Arabic sources. Over a four-month period, these eighteen sources will include a very large number of stories. We anticipate using about 61,000 stories after sub-sampling. LDC collected data from each source once per day on which the source was available.

To reduce costs associated with licensing and transcription, LDC has sub-sampled the data to select more than half of the days that a source is available per week and to stagger sources so that there are multiple sources per day per language. While collecting the material LDC managed intellectual property negotiations necessary to make this data available for research use in the TDT project and beyond. We also identified external transcription bureaus, negotiated rates and manage the offsite transcription of this material. Transcription quality is similar to that available from commercial closed captioning. No attempt is made to produce transcripts of the quality used in speech engineering research projects such as Hub-4.

With transcripts in hand, LDC segments the text to identify individual stories. The transcription bureaus provide the first-pass story segmentation and LDC annotators perform the second pass during which they will listen to the audio of the entire broadcast while viewing the corresponding waveform display and text intermediary and add, remove or re-position story boundaries. Annotators also classify each boundary as beginning a 1) news story 2) miscellaneous text section or 3) untranscribed section. As a quality check, we use patterns of segment boundaries observed in previous corpora to guide segmentation.

For thee months of broadcast news, LDC typically defines 60 topics by a random process that gives each month of data from each source an equal opportunity to contribute a topic. To improve consistency, we perform research on each topic before annotation begins and to maximize annotator efficiency, we use a search guided annotation procedure developed in 2000.

LDC annotates all English, Mandarin and Arabic stories sampled against all topics defined. The first pass involves submitting the concatenation of all on-topic stories as a query into the corpus. During this first stage, on-topic stories include the seed story itself and any stories found during topic definition and research. The annotators then read through the stories in the relevance ranked list until reaching the "off-topic threshold" defined as a 2:1 ration of off-topic to on-topic stories in which the last 10 stories read were off-topic. In the second stage, annotators will iterate their searches using the concatenation of all on-topic stories as they continued to find them. During the third stage, annotators issue new queries using the text of the topic research document and topic explications. As before, annotators will read the relevance ranked list of returns to reach the off-topic threshold before progressing to the next stage. In the fourth stage, annotators will think creatively to conduct additional manual searches through the corpus.

Corpus creation procedures for TREC differ in several ways. Where TDT topics are based upon a seminal event reported in the news, TREC topics are broader and answer a general question about events in the news. Additionally TREC does not attempt an exhaustive annotation of all stories in a corpus prior to technology evaluation. Instead, TREC research sites submit all of the stories they believe to be on-topic. Human assessors then review the returns compiled from all sites that have the highest probability of being on-topic. LDC performed topic definition and system assessment for TREC's Arabic-English cross-lingual track in 2001 and will repeat the annotation and assessment with a larger set of topics in 2002. The Arabic corpus contained over 383,000 stories from 6 years of Agence France Presse newswire. Systems were assessed for 25 topics.

10. Evaluation Resources, Extraction

The DARPA program in Automatic Content Extraction has developed specifications for the annotation of all entities (person, locations, organizations, etc) and relations mentioned in a text. These specifications are clearly relevant to TIDES sponsored work in information extraction. LDC has joined three other sites in annotating newswire, newspaper text and broadcast news transcripts for entities and relations. The corpora are based on raw data from broadcast news, newswire and stories clipped from printed newspapers. Much of the ACE data can also

be found in the TDT corpora. This overlap is useful and will be encouraged throughout TIDES as well.

The types of annotation specified for ACE involve identifying entities (persons, locations, organizations, etc) in text including the maximal extent of the string that represents the entity and its type. Because ACE annotators also mark co-reference and metonymy, they are building up a database of all forms in which entities are mentioned in a selected set of texts. Recently some groups began annotating relations among entities while other are beginning the annotation in Chinese. ACE annotation use MITRE's Alembic Workbench. Starting in FY2002, ACE annotation work is carried out under the TIDES umbrella. IN 2002, LDC is the only site performing entity annotation in English. LDC has also begun relation annotation in English and entity annotation in Chinese

11. Evaluation Resources, Summarization

Research in information summarization may rely upon two kinds of data 1) collections of source documents that have been annotated to indicate sections that express information that is new, or deemed important or that bears on the topicality of the document 2) collection of full-length document and separate summaries of those documents. Within the TIDES community, the data specification for summarization research is still under development.

The Johns Hopkins summer workshop in speech and language engineering experimented with one possible approach, LDC created 40 topics from the Hong Kong News Corpus. For each topic, annotators labeled 100 documents for relevance. For ten documents per topic, annotators also labeled each sentence in the document for its "importance" to the topic as a whole; sentence-labeling was repeated by three annotators per topic. Query Formulation was based upon clusters generated automatically. LDC used story clusters to identify potentially useful topics. Annotators read documents within each cluster. Any cluster that did not discuss an appropriate query was discarded. For any good cluster, annotators removed those documents that were relevant. A topic title was then created based on the content of the documents in the cluster.

For each good topic, annotators performed document-based annotation. Annotators use the EZQuery search engine to create a list of 100 relevance-ranked documents based on a query of all stories previously known to be on-topic. Annotators then read and labeled each story in this list for relevance, using the labels YES (relevant), BRIEF (less than 10% relevant) or NO (irrelevant).

For each of the topics labeled during the second phase of annotation, the annotator who conducted document relevance assessment for that query completed sentence-based relevance judgments for the 10 documents the search engine gave the highest ranking for that topic. In addition to the original annotator, two additional annotators, working independently, completed sentence-based relevance for each of the queries. LDC will release this annotated data in 2002.

12. Evaluation Material, Machine Translation

The TIDES machine translation has until recently focused on formulating a system evaluation metrics. To support this work, LDC has produced multiple clusters of 10K words that have each been translated by a number of translators running the gamut from low to high quality. LDC selected documents for these experiments, acquired distribution rights, outsourced Chinese-English and Arabic-English human translations and ran commercial MT system to produce automatic translations. Source material is sampled from existing corpora to provide overlap with other communities and from new material so that sites can test the generality of their algorithms. LDC has remained in close contact with the MT research community so that the resources we will provide will continue to match its evolving needs.

Our guidelines for this project were based upon the guidelines developed for the translation of the Chinese Treebank with two important modifications. First, we instructed agencies to describe the translation teams used. The Chinese Treebank guidelines emphasize "faithfulness" in terms of vocabulary, syntactic structures, and of course semantic contents providing specific examples. This, coupled with the relatively constrained writing in the original documents, threatens to produce less variation than desired. The Chinese Treebank guidelines were intended for a single translation of a very precious data set and require uniformly excellent quality. We relax the "faithfulness" constraint in our translations

In this project, quality assurance means making sure the translation agencies understand the task and checking returning translations for coverage and format. The LDC does not in any way modify the content of the translation because variation in quality is desired.

13. Intellectual Property Arrangements

Much of the material described above is based upon large volumes of text and speech best collected from commercial providers. Commercial sources may require the negotiation of agreements that permit the distribution of data to researchers while constraining the use of the material to linguistic education, research and technology development. LDC has negotiated such agreements since 1992. Other arrangements are possible and warrant consideration. Some material may be in the public domain. Some uses of material may fit under the doctrine of Fair Use. Furthermore, Copyright Law is dynamic and responds to changes in the information technology. LDC endeavors to keep abreast of changes in copyright law as they affect information distribution. In any case, LDC coordinates all necessary intellectual property arrangements for multiple research programs including TIDES to make resources gathered in this way available to the broader research communities.

14. Resource Distribution Infrastructure

Researchers in speech and language technologies have realized for at least three decades that shared linguistic resources are an important stimulant to progress. DARPA sponsored common task research programs rely heavily upon shared resources. LDC was in fact created specifically to facilitate research sharing. With the support of the TIDES sponsors, LDC has extended its role in research sharing by coordinating all resource acquisition for TIDES. We focus our efforts on resources that can be shared at least as broadly as the sponsoring program but

more typically to all communities working in linguistic education, research and technology development.

15. References

ACE, 2000, Automatic Content Extraction [www.nist.gov/speech/tests/ace].

Bird, Steven, Kazuaki Maeda, Xiaoyi Ma, Haejoong Lee, 2002, MultiTrans and TableTrans: Annotation Tools Based on the Annotation Graph Toolkit (AGTK), Proceedings of the Third International Language Resources and Evaluation Conference, Las Palmas, Spain, May-June 2002.

Bird, Steven, Mark Liberman, 2002, A Call for Open Source Lexicons, Proceedings of the Third International Language Resources and Evaluation Conference, Las Palmas, Spain, May-June 2002.

Bird, Steven, Hans Uskoreit, Gary Simons, 2002, The Open Language Archives Community, Proceedings of the Third International Language Resources and Evaluation Conference, Las Palmas, Spain, May-June 2002.

Bird, Steven and Mark Liberman, 1999, Linguistic Annotation Page, [www.ldc.upenn.edu/annotation]

Cieri, Christopher, Dave Graff, Mark Liberman, Nii Martey and Stephanie Strassel, 2000, Large Multilingual Broadcast News Corpora for Cooperative Research in Topic Detection and Tracking: The TDT2 and TDT3 Corpus Efforts, In Proceedings of the Second International Language Resources and Evaluation Conference, Athens, Greece, May 2000.

Doddington, G. (1999). The 1999 Topic Detection and Tracking (TDT) Task Definition and Evaluation Plan. Available at http://www.nist.gov/TDT.

Lamel, Lori, Fabrice Lefevre, Jean-Luc Gauvain and Gilles Adda, 2001, Portability Issues for Speech Recognition Technologies, HLT 2001: Proceedings of the First International Conference on Human Language Technology Research, San Diego, CA, March 18-21, 2001.

LDC, 2000, Linguistic Data Consortium Homepage [http://www.ldc.upenn.edu]

Ma, Xiaoyi and Mark Liberman, 1999, BITS: A Method for Bilingual Text Search over the Web, presented at Machine Translation Summit VII, September 13th, 1999, Kent Ridge Digital Labs, National University of Singapore, [www.ldc.upenn.edu/Papers/MTSVII1999/BITS.ps]

Papineni, Kishore, Salim Roukos, Todd, Ward, John Henderson, Florence Reeder, 2002, Corpus-Based Comprehensive and Diagnostic MT Evaluation: Initial Arabic, Chinese, French and Spanish Results, HLT 2002: Proceedings of the Second International Conference on Human Language Technology Research, San Diego, CA, March 24-27, 2002.

Resnik, Philip, 1999, Mining the Web for Bilingual Text, ACL 1999: 37th Annual Meeting of the Association for Computational Linguistics (ACL' 99), College Park, Maryland, June 1999.

Saggion, Horacio, Dragomir Radev, Simone Teufel, Wai Lam and Stephanie Strassel, 2002, Developing Infrastructure for the Evaluation of Single and Multi-document Summarization Systems in a Multi-lingual Environment, Proceedings of the Third International Language Resources and Evaluation Conference, Las Palmas, Spain, May-June 2002.

Strassel, Stephanie and Christopher Cieri, 2002, Resource Development for Topic Detection and Tracking Research: The TDT-4 Corpus, Proceedings of the Third International Language Resources and Evaluation Conference, Las Palmas, Spain, May-June 2002.

TIDES, 2000, DARPA Program in Translingual Information Detection Extraction and Summarization [www.arpa.mil/ito/research/tides]

Wayne, Charles, 2000, Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation, In Proceedings of the Second International Language Resources and Evaluation Conference, Athens, Greece, May 2000.

Wayne, Charles, 1998, Topic Detection & Tracking: A Case Study in Corpus Creation & Evaluation Methodologies, In Proceedings of Language Resources and Evaluation Conference, Granada, Spain, May 1998.