Mining the Web for Bilingual Text

Philip Resnik*
Dept. of Linguistics/Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742
resnik@umiacs.umd.edu

Abstract

STRAND (Resnik, 1998) is a language-independent system for automatic discovery of text in parallel translation on the World Wide Web. This paper extends the preliminary STRAND results by adding automatic language identification, scaling up by orders of magnitude, and formally evaluating performance. The most recent end-product is an automatically acquired parallel corpus comprising 2491 English-French document pairs, approximately 1.5 million words per language.

1 Introduction

Text in parallel translation is a valuable resource in natural language processing. Statistical methods in machine translation (e.g. (Brown et al., 1990)) typically rely on large quantities of bilingual text aligned at the document or sentence level, and a number of approaches in the burgeoning field of cross-language information retrieval exploit parallel corpora either in place of or in addition to mappings between languages based on information from bilingual dictionaries (Davis and Dunning, 1995; Landauer and Littman, 1990; Hull and Oard, 1997; Oard, 1997). Despite the utility of such data, however, sources of bilingual text are subject to such limitations as licensing restrictions, usage fees, restricted domains or genres, and dated text (such as 1980's Canadian politics); or such sources simply may not exist for language pairs of interest.

Although the majority of Web content is in English, it also shows great promise as a source of multilingual content. Using figures from the Babel survey of multilinguality on the Web (http://www.isoc.org/), it is possible to estimate that as of June, 1997, there were on the order of 63000 primarily non-English Web servers, ranging over 14 languages. Moreover, a follow-up investigation of the non-English servers suggests that nearly a third contain some useful cross-language data, such as parallel English on the page or links to parallel English pages — the follow-up also found pages in five languages not identified by the Babel study (Catalan, Chinese, Hungarian, Icelandic, and Arabic; Michael Littman, personal communication). Given the continued explosive increase in the size of the Web, the trend toward business organizations that cross national boundaries, and high levels of competition for consumers in a global marketplace, it seems impossible not to view multilingual content on the Web as an expanding resource. Moreover, it is a dynamic resource, changing in content as the world changes. For example, Diekema et al., in a presentation at the 1998 TREC-7 conference (Voorhees and Harman, 1998), observed that the performance of their cross-language information retrieval was hurt by lexical gaps such as *Bosnia*/*Bosnie* — this illustrates a highly topical missing pair in their static lexical resource (which was based on WordNet 1.5). And Gey et al., also at TREC-7, observed that in doing cross-language retrieval using commercial machine translation systems, gaps in the lexicon (their example was *acupuncture*/*Akupunktur*) could make the difference between precision of 0.08 and precision of 0.83 on individual queries.

Resnik (1998) presented an algorithm called

* This work was supported by Department of Defense contract MDA90496C1250, DARPA/ITO Contract N66001-97-C-8540, and a research grant from Sun Microsystems Laboratories. The author gratefully acknowledges the comments of the anonymous reviewers, helpful discussions with Dan Melamed and Doug Oard, and the assistance of Jeff Allen in the French-English experimental evaluation.

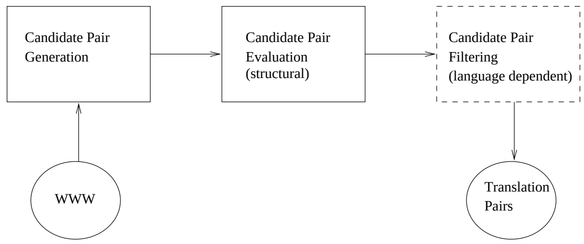

Figure 1: The STRAND architecture

STRAND (Structural Translation Recognition for Acquiring Natural Data) designed to explore the Web as a source of parallel text, demonstrating its potential with a small-scale evaluation based on the author's judgments. After briefly reviewing the STRAND architecture and preliminary results (Section 2), this paper goes beyond that preliminary work in two significant ways. First, the framework is extended to include a filtering stage that uses automatic language identification to eliminate an important class of false positives: documents that appear structurally to be parallel translations but are in fact not in the languages of interest. The system is then run on a somewhat larger scale and evaluated formally for English and Spanish using measures of agreement with independent human judges, precision, and recall (Section 3). Second, the algorithm is scaled up more seriously to generate large numbers of parallel documents, this time for English and French, and again subjected to formal evaluation (Section 4). The concrete end result reported here is an automatically acquired English-French parallel corpus of Web documents comprising 2491 document pairs, approximately 1.5 million words per language (without markup), containing little or no noise.

2 STRAND Preliminaries

This section is a brief summary of the STRAND system and previously reported preliminary results (Resnik, 1998).

The STRAND architecture is organized as a pipeline, beginning with a *candidate generation* stage that (over-)generates candidate pairs of documents that might be parallel translations. (See Figure 1.) The first implementation of the generation stage used a query to the Altavista search engine to generate pages that could be viewed as "parents" of pages in parallel translation, by asking for pages containing one portion of anchor text (the readable material in a hyperlink) containing the string "English" within a fixed distance of another anchor text containing the string "Spanish". (The matching process was case-insensitive.) This generated many good pairs of pages, such as those pointed to by hyperlinks reading *Click here for English version* and *Click here for Spanish version*, as well as many bad pairs, such as university pages containing links to *English Literature* in close proximity to *Spanish Literature*.

The candidate generation stage is followed by a *candidate evaluation* stage that represents the core of the approach, filtering out bad candidates from the set of generated page pairs. It employs a structural recognition algorithm exploiting the fact that Web pages in parallel translation are invariably very similar in the way they are structured — hence the 's' in STRAND. For example, see Figure 2.

The structural recognition algorithm first runs both documents through a transducer that reduces each to a linear sequence of tokens corresponding to HTML markup elements, interspersed with tokens representing undifferentiated "chunks" of text. For example, the transducer would replace the HTML source text `<TITLE>ACL'99 Conference Home Page</TITLE>` with the three tokens `[BEGIN:TITLE]`, `[Chunk:24]`, and `[END:TITLE]`. The number inside the chunk token is the length of the text chunk, not counting whitespace; from this point on only the length of the text chunks is used, and therefore the structural filtering algorithm is completely language independent.

Given the transducer's output for each document, the structural filtering stage aligns the two streams of tokens by applying a standard, widely available dynamic programming algorithm for finding an optimal alignment between two linear sequences.[1] This alignment matches identical markup tokens to each other as much as possible, identifies runs of unmatched tokens that appear to exist only in one sequence but not the other, and marks pairs of non-identical tokens that were forced to be matched to each other in order to obtain the best alignment pos-

[1] Known to many programmers as `diff`.

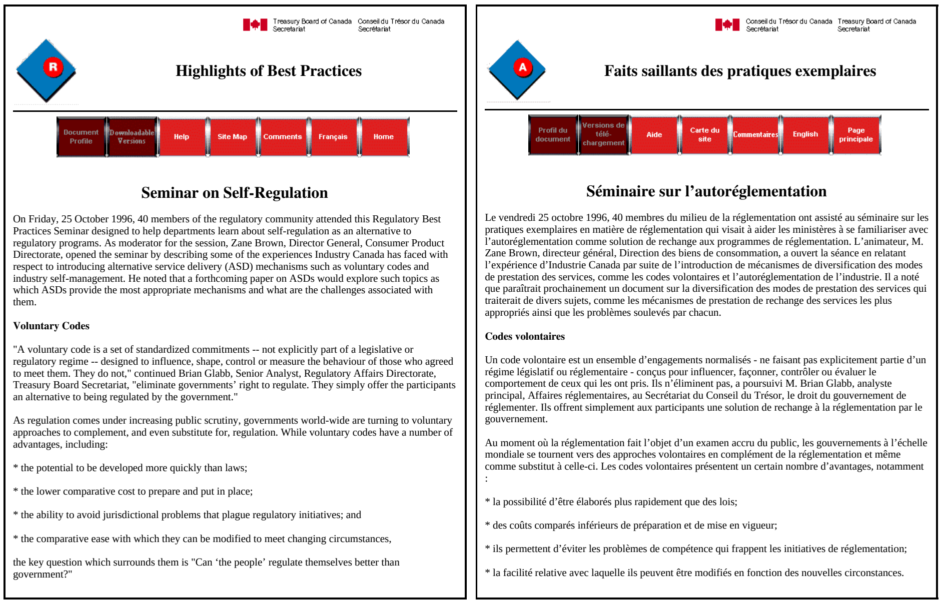

Figure 2: Structural similarity in parallel translations on the Web

sible.[2] At this point, if there were too many unmatched tokens, the candidate pair is taken to be *prima facie* unacceptable and immediately filtered out.

Otherwise, the algorithm extracts from the alignment those pairs of chunk tokens that were matched to each other in order to obtain the best alignments.[3] It then computes the correlation between the lengths of these non-markup text chunks. As is well known, there is a reliably linear relationship in the lengths of text translations — small pieces of source text translate to small pieces of target text, medium to medium, and large to large. Therefore we can apply a standard statistical hypothesis test, and if $p < .05$ we can conclude that the lengths are reliably correlated and accept the page pair as likely to be translations of each other. Otherwise, this candidate page pair is filtered out.[4]

In the preliminary evaluation, I generated a test set containing 90 English-Spanish candidate pairs, using the candidate generation stage as just described. I evaluated these candidates by hand, identifying 24 as true translation pairs.[5] Of these 24, STRAND identified 15 as true translation pairs, for a recall of 62.5%. Perhaps more important, it only generated 2 additional translation pairs incorrectly, for a precision of $15/17 = 88.2\%$.

3 Adding Language Identification

In the original STRAND architecture, additional filtering stages were envisaged as possible (see Figure 1), including such language-dependent processes as automatic language identification and content-based comparison of structually aligned document segments using cognate matching or existing bilingual dictionaries. Such stages were initially avoided in order to keep the system simple, lightweight, and independent of linguistic resources. How-

[2] An anonymous reviewer observes that `diff` has no preference for aligning chunks of similar lengths, which in some cases might lead to a poor alignment when a good one exists. This could result in a failure to identify true translations and is worth investigating further.

[3] Chunk tokens with *exactly* equal lengths are excluded; see (Resnik, 1998) for reasons and other details of the algorithm.

[4] The level of significance ($p < .05$) was the initial selection during algorithm development, and never changed. This, the unmatched-tokens threshold for *prima facie* rejection due to mismatches (20%), and the maximum distance between hyperlinks in the generation stage (10 lines), are parameters of the algorithm that were determined during development using a small amount of arbitrarily selected French-English data downloaded from the Web. These values work well in practice and have not been varied systematically; their values were fixed in advance of the preliminary evaluation and have not been changed since.

[5] The complete test set and my judgments for this preliminary evaluation can be found at `http://umiacs.umd.edu/~resnik/amta98/`.

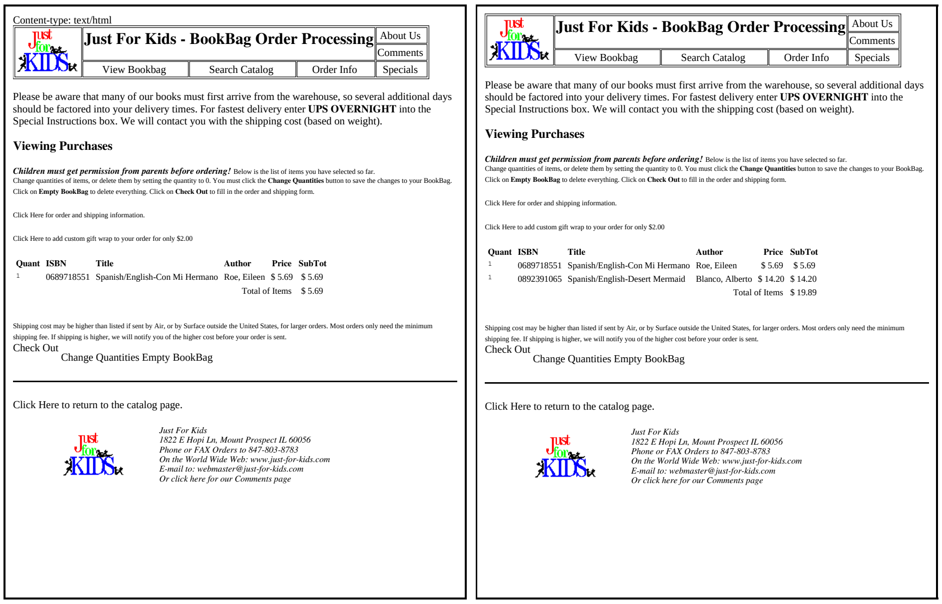

Figure 3: Structurally similar pages that are not translations

ever, in conducting an error analysis for the preliminary evaluation, and further exploring the characteristics of parallel Web pages, it became evident that such processing would be important in addressing one large class of potential false positives. Figure 3 illustrates: it shows two documents that are generated by looking for "parent" pages containing hyperlinks to *English* and *Spanish*, which pass the structural filter with flying colors. The problem is potentially acute if the generation stage happens to yield up many pairs of pages that come from on-line catalogues or other Web sites having large numbers of pages with a conventional structure.

There is, of course, an obvious solution that will handle most such cases: making sure that the two pages are actually written in the languages they are supposed to be written in. In order to filter out candidate page pairs that fail this test, statistical language identification based on character n-grams was added to the system (Dunning, 1994). Although this does introduce a need for language-specific training data for the two languages under consideration, it is a very mild form of language dependence: Dunning and others have shown that when classifying strings on the order of hundreds or thousands of characters, which is typical of the non-markup text in Web pages, it is possible to discriminate languages with accuracy in the high 90% range for many or most language pairs given as little as 50k characters per language as training material.

For the language filtering stage of STRAND, the following criterion was adopted: given two documents d_1 and d_2 that are supposed to be in languages L_1 and L_2, keep the document pair iff $\Pr(L_1|d_1) > \Pr(L_2|d_1)$ and $\Pr(L_2|d_2) > \Pr(L_1|d_2)$. For English and Spanish, this translates as a simple requirement that the "English" page look more like English than Spanish, and that the "Spanish" page look more like Spanish than English. Language identification is performed on the plain-text versions of the pages. Character 5-gram models for languages under consideration are constructed using 100k characters of training data from the European Corpus Initiative (ECI), available from the Linguistic Data Consortium (LDC).

In a formal evaluation, STRAND with the new language identification stage was run for English and Spanish, starting from the top 1000 hits yielded up by Altavista in the candidate generation stage, leading to a set of 913 candidate pairs. A test set of 179 items was generated for annotation by human judges, containing:

- All the pairs marked GOOD (i.e. translations) by STRAND (61); these are the pairs that passed both the structural and language identification filter.
- All the pairs filtered out via language iden-

tification (73)

- A random sample of the pairs filtered out structurally (45)

It was impractical to manually evaluate all pairs filtered out structurally, owing to the time required for judgments and the desire for two independent judgments per pair in order to assess inter-judge reliability.

The two judges were both native speakers of Spanish with high proficiency in English, neither previously familiar with the project. They worked independently, using a Web browser to access test pairs in a fashion that allowed them to view pairs side by side. The judges were told they were helping to evaluate a system that identifies pages on the Web that are translations of each other, and were instructed to make decisions according to the following criterion:

> *Is this pair of pages intended to show the same material to two different users, one a reader of English and the other a reader of Spanish?*

The phrasing of the criterion required some consideration, since in previous experience with human judges and translations I have found that judges are frequently unhappy with the quality of the translations they are looking at. For present purposes it was required neither that the document pair represent a perfect translation (whatever that might be), nor even necessarily a good one: STRAND was being tested not on its ability to determine translation *quality*, which might or might not be a criterion for inclusion in a parallel corpus, but rather its ability to facilitate the task of locating page pairs that one might reasonably include in a corpus undifferentiated by quality (or potentially post-filtered manually).

The judges were permitted three responses:

- Yes: translations of each other
- No: not translations of each other
- Unable to tell

When computing evaluation measures, page pairs classified in the third category by a human judge, for whatever reason, were excluded from consideration.

Comparison	N	Pr(Agree)	κ
J1, J2:	106	0.85	0.70
J1, STRAND:	165	0.91	0.79
J2, STRAND:	113	0.81	0.61
J1 $\cap$ J2, STRAND:	90	0.91	0.82

Table 1: English-Spanish evaluation

Table 1 shows agreement measures between the two judges, between STRAND and each individual judge, and the agreement between STRAND and the intersection of the two judges' annotations — that is, STRAND evaluated against only those cases where the two judges agreed, which are therefore the items we can regard with the highest confidence. The table also shows Cohen's κ, an agreement measure that corrects for chance agreement (Carletta, 1996); the most important κ value in the table is the value of 0.7 for the two human judges, which can be interpreted as sufficiently high to indicate that the task is reasonably well defined. (As a rule of thumb, classification tasks with $\kappa < 0.6$ are generally thought of as suspect in this regard.) The value of N is the number of pairs that were included, after excluding those for which the human judgement in the comparison was undecided.

Since the cases where the two judges agreed can be considered the most reliable, these were used as the basis for the computation of recall and precision. For this reason, and because the human-judged set included only a sample of the full set evaluated by STRAND, it was necessary to extrapolate from the judged (by both judges) set to the full set in order to compute recall/precision figures; hence these figures are reported as estimates. Precision is estimated as the proportion of pages judged GOOD by STRAND that were also judged to be good (i.e. "yes") by both judges — this figure is 92.1% Recall is estimated as the number of pairs that *should* have been judged GOOD by STRAND (i.e. that recieved a "yes" from both judges) that STRAND indeed marked GOOD — this figure is 47.3%.

These results can be read as saying that of every 10 document pairs included by STRAND in a parallel corpus acquired fully automatically from the Web, fewer than 1 pair on average was included in error. Equivalently, one could say that the resulting corpus contains only about

8% noise. Moreover, at least for the confidently judged cases, STRAND is in agreement with the combined human judgment more often than the human judges agree with each other. The recall figure indicates that for every true translation pair it accepts, STRAND must also incorrectly reject a true translation pair. Alternatively, this can be interpreted as saying that the filtering process has the system identifying about half of the pairs it could in principle have found given the candidates produced by the generation stage. Error analysis suggests that recall could be increased (at a possible cost to precision) by making structural filtering more intelligent; for example, ignoring some types of markup (such as italics) when computing alignments. However, I presume that if the number M of translation pairs on the Web is large, then half of M is also large. Therefore I focus on increasing the total yield by attempting to bring the number of generated candidate pairs closer to M, as described in the next section.

4 Scaling Up Candidate Generation

The preliminary experiments and the new experiment reported in the previous section made use of the Altavista search engine to locate "parent" pages, pointing off to multiple language versions of the same text. However, the same basic mechanism is easily extended to locate "sibling" pages: cases where the page in one language contains a link directly to the translated page in the other language. Exploration of the Web suggests that parent pages and sibling pages cover the major relationships between parallel translations on the Web. Some sites with bilingual text are arranged according to a third principle: they contain a completely separate monolingual sub-tree for each language, with only the single top-level home page pointing off to the root page of single-language version of the site. As a first step in increasing the number of generated candidate page pairs, STRAND was extended to permit both parent and sibling search criteria. Relating monolingual sub-trees is an issue for future work.

In principle, using Altavista queries for the candidate generation stage should enable STRAND to locate every page pair in the Altavista index that meets the search criteria. This likely to be an upper bound on the candidates that can be obtained without building a Web crawler dedicated to the task, since one of Altavista's distinguishing features is the size of its index. In practice, however, the user interface for Altavista appears to limit the number of hits returned to about the first 1000. It was possible to break this barrier by using a feature of Altavista's "Advanced Search": including a range of dates in a query's selection criteria. Having already redesigned the STRAND generation component to permit multiple queries (in order to allow search for both parent and sibling pages), each query in the query set was transformed into a set of mutually exclusive queries based on a one-day range; for example, one version of a query would restrict the result to pages last updated on 30 November 1998, the next 29 November 1998, and so forth.

Comparison	N	Pr(Agree)	κ
J1, J2:	267	0.98	0.95
J1, STRAND:	273	0.84	0.65
J2, STRAND:	315	0.85	0.63
J1 $\cap$ J2, STRAND:	261	0.86	0.68

Table 2: English-French evaluation

Although the solution granularity was not perfect — searches for some days still bumped up against the 1000-hit maximum — use of both parent and sibling queries with date-range restricted queries increased the productivity of the candidate generation component by an order of magnitude. The scaled-up system was run for English-French document pairs in late November, 1998, and the generation component produced 16763 candidate page pairs (with duplicates removed), an 18-fold increase over the previous experiment. After eliminating 3153 page pairs that were either exact duplicates or irretrievable, STRAND's structural filtering removed 9820 candidate page pairs, and the language identification component removed another 414. The remaining pairs identified as GOOD — i.e. those that STRAND considered to be parallel translations — comprise a parallel corpus of 3376 document pairs.

A formal evaluation, conducted in the same fashion as the previous experiment, yields the agreement data in Table 2. Using the cases where the two human judgments agree as ground truth, precision of the system is estimated at 79.5%, and recall at 70.3%.

Comparison	N	Pr(Agree)	κ
J1, J2:	267	0.98	0.95
J1, STRAND:	273	0.88	0.70
J2, STRAND:	315	0.88	0.69
J1 $\cap$ J2, STRAND:	261	0.90	0.75

Table 3: English-French evaluation with stricter language ID criterion

A look at STRAND's errors quickly identifies the major source of error as a shortcoming of the language identification module: its implicit assumption that every document is either in English or in French. This assumption was violated by a set of candidates in the test set, all from the same site, that pair *Dutch* pages with French. The language identification criterion adopted in the previous section requires only that the Dutch pages look *more like English* than like French, which in most cases is true. This problem is easily resolved by training the existing language identification component with a wider range of languages, and then adopting a stricter filtering criterion requiring that $\Pr(\text{English}|d_1) > \Pr(\text{L}|d_1)$ for *every* language L in that range, and that d_2 meet the corresponding requirement for French.[6] Doing so leads to the results in Table 3.

This translates into an estimated 100% precision against 64.1% recall, with a yield of 2491 documents, approximately 1.5 million words per language as counted after removal of HTML markup. That is, with a reasonable though admittedly post-hoc revision of the language identification criterion, comparison with human subjects suggests the acquired corpus is non-trivial and essentially noise free, and moreover, that the system excludes only a third of the pages that should have been kept. Naturally this will need to be verified in a new evaluation on fresh data.

[6] Language ID across a wide range of languages is not difficult to obtain. E.g. see the 13-language set of the freely available CMU stochastic language identifier (http://www.cs.cmu.edu/~dougb/ident.html), the 18-language set of the Sun Language ID Engine (http://www.sunlabs.com/research/ila/demo/index.html), or the 31-language set of the XRCE Language Identifier (http://www.rxrc.xerox.com/research/mltt/Tools/guesser.html). Here I used the language ID method of the previous section trained with profiles of Danish, Dutch, English, French, German, Italian, Norwegian, Portuguese, Spanish, and Swedish.

5 Conclusions

This paper places acquisition of parallel text from the Web on solid empirical footing, making a number of contributions that go beyond the preliminary study. The system has been extended with automated language identification, and scaled up to the point where a non-trivial parallel corpus of English and French can be produced completely automatically from the World Wide Web. In the process, it was discovered that the most lightweight use of language identification, restricted to just the the language pair of interest, needed to be revised in favor of a strategy that includes identification over a wide range of languages. Rigorous evaluation using human judges suggests that the technique produces an extremely clean corpus — noise estimated at between 0 and 8% — even without human intervention, requiring no more resources per language than a relatively small sample of text used to train automatic language identification.

Two directions for future work are apparent. First, experiments need to be done using languages that are less common on the Web. Likely first pairs to try include English-Korean, English-Italian, and English-Greek. Inspection of Web sites — those with bilingual text identified by STRAND and those without — suggests that the strategy of using Altavista to generate candidate pairs could be improved upon significantly by adding a true Web crawler to "mine" sites where bilingual text is known to be available, e.g. sites uncovered by a first pass of the system using the Altavista engine. I would conjecture that for English-French there is an order of magnitude more bilingual text on the Web than that uncovered in this early stage of research.

A second natural direction is the application of Web-based parallel text in applications such as lexical acquisition and cross-language information retrieval — especially since a side-effect of the core STRAND algorithm is aligned "chunks", i.e. non-markup segments found to correspond to each other based on alignment of the markup. Preliminary experiments using even small amounts of these data suggest that standard techniques, such as cross-language lexical association, can uncover useful data.

References

P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, R. Mercer, and P. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Jean Carletta. 1996. Assessing agreement on classification tasks: the Kappa statistic. *Computational Linguistics*, 22(2):249–254, June.

Mark Davis and Ted Dunning. 1995. A TREC evaluation of query translation methods for multi-lingual text retrieval. In *Fourth Text Retrieval Conference (TREC-4)*. NIST.

Ted Dunning. 1994. Statistical identification of language. Computing Research Laboratory Technical Memo MCCS 94-273, New Mexico State University, Las Cruces, New Mexico.

David A. Hull and Douglas W. Oard. 1997. Symposium on cross-language text and speech retrieval. Technical Report SS-97-04, American Association for Artificial Intelligence, Menlo Park, CA, March.

Thomas K. Landauer and Michael L. Littman. 1990. Fully automatic cross-language document retrieval using latent semantic indexing. In *Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research*, pages pages 31–38, UW Centre for the New OED and Text Research, Waterloo, Ontario, October.

Douglas W. Oard. 1997. Cross-language text retrieval research in the USA. In *Third DELOS Workshop*. European Research Consortium for Informatics and Mathematics March.

Philip Resnik. 1998. Parallel strands: A preliminary investigation into mining the web for bilingual text. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas, AMTA-98, in Lecture Notes in Artificial Intelligence, 1529*, Langhorne, PA, October 28-31.

E. M. Voorhees and D. K. Harman. 1998. The seventh Text REtrieval Conference (TREC-7). NIST special publication, Gaithersburg, Maryland, November 9–11. `http://trec.nist.gov/pubs.html`.