Multi-lingual Text Filtering Using Semantic Modeling

James R. Driscoll*

and

Sara Abbott, Kai-Lin Hu, Michael Miller, Gary Theis

Department of Computer Science
University of Central Florida
Orlando, Florida 32816

Abstract

Semantic Modeling is used to investigate multilingual text filtering. In our approach, the Entity-Relationship (ER) Model is used as a basis for descriptions of information preferences (profiles) in the information filtering process. A profile is viewed as having both a static aspect and a dynamic aspect. The static aspect of a profile can be represented as an ER schema; and the dynamic aspect of the profile can be represented by synonyms of schema components and domain values for schema attributes. For TREC-4, the routing task and the Spanish adhoc task are accomplished using this technique. For the routing task, a large amount of time was spent in an effort to optimize filter performance using the training data that was available for the routing topics. For the Spanish adhoc task, a large amount of time was spent using external sources to develop good filters; in addition, some time was spent implementing a program to help port our approach to this second language. A multi-lingual (English, French, German, and Spanish) experiment is also reported.

Introduction - The Filtering Task and Profiles

Our approach to filtering was first presented at TREC-3 [4]. The filtering process is based on descriptions of individual (or group) information preferences, often called profiles. Profiles typically represent long-term interests. Information filtering is concerned with repeated uses of the profiles, and the profile is assumed to be a correct specification of an information interest [2].

We view a profile as having both a static aspect and a dynamic aspect. We present a procedure for representing the user need statement of a TREC topic as a database Entity-Relationship (ER) schema. An ER schema becomes the static aspect of a profile. For a schema, a synonym list is created for each of the schema components, and a domain value list is created for each of the schema attributes. These lists become the major part of the dynamic aspect of a profile. Our filtering procedure uses the dynamic aspect of a profile to detect relevant documents.

TREC topics are descriptions of information to be considered relevant, and these must be transformed into filter profiles. Each TREC topic is in the form of a highly-formatted, natural language, user need statement. Refer to Figure 1 for an example. This is TREC Topic 173 which concerns smoking bans.

A discussion of semantic modeling and our procedure for making a profile which represents a TREC topic appears in [4]. In this paper, we explain our approach using our TREC-4 filter for Topic 173 as an example.

* Dr. Driscoll's current affiliation is Praxis Technologies, 280 West Canton Avenue, Suite 230, Winter Park, Florida, 32789. Phone: (407) 647-0001, Fax: (407) 628-1832, E-mail: driscoll@prx.com or jdriscol@prx.com.

<top>

<num> Number: 173

<title> Topic: Smoking Bans

<desc> Description:

Document will provide data on smoking bans initiated worldwide in the public and private sector workplace, on various modes of public transportation, and in commercial advertising.

<narr> Narrative:

A relevant document would include data on smoking bans that have been initiated worldwide in the workplace, on various modes of transportation, and in commercial advertising. Relevant information would include such data as who initiated the ban, affected areas, enforcement policy/procedures if any, and any penalties that may be imposed. Also relevant would be tobacco company reactions, legislation imposing a smoking ban, and legal actions or court decisions pertaining to a smoking ban. Not relevant would be documents containing public or private sector comments on smoking in general but not related to a specific ban. Also not relevant would be documents related to health hazards associated with smoking and not referring to a smoking ban.

</top>

Figure 1. TREC Topic 173.

Details and an Example

In our approach, a filter for a TREC topic includes a series of lists (files). A list is either a synonym list or a domain list. Refer to [4] for the theory behind synonym and domain lists. A synonym list is just a list of words or phrases that indicate the same concept. For example, synonyms for the word "ban". Sometimes, several synonym lists can be merged into one list. We also include the various forms of a word in synonym lists. A domain list is just a list of words or phrases that represent the various allowed values for a particular item. For example, the names of public places could be banks, department stores, elevators, hospitals, etc. Sometimes, the values that should be in a domain list are not known, or only partially known.

Our approach allows weights to be assigned to indicate the importance of each list. We also make use of a window of text when text is examined. Our system has the ability to count "hits" in a window for various combinations of the lists. The window size, the file combinations, and file weights can be adjusted.

At the present time we manually create filter profiles. We determine the significant domain and synonym lists from a study of the TREC topic. Then we initially populate the lists using thesauri, dictionaries, and whatever other reference sources we can find; sometimes a list remains empty. Populating the lists in this manner is the procedure we used to build filters for the Spanish adhoc task. For the routing task, we used training data by reading relevant and not-relevant documents to build filters. Finally, we create an information (INF) file to specify the window size, and the various file combinations and file weights.

To establish a filter for a TREC topic, and do TREC filtering experiments, we have a standard text scanning program which inputs the window size, the domain and synonym files, and one or more variations (which also indicate weights) of the lists. The scanning program then moves across TREC document collections, producing a ranked list of relevant documents. We have used the TREC training data to modify the dynamic aspect of a profile. This is accomplished by using viewed relevant and non-relevant documents to adjust the window size and make additions and deletions to the domain and synonym lists. We have developed a few utility programs to help us do this quickly.

In its current implementation, the scanner requires a stream of text delimited by standard SGML. The only specific markers actually used are the <DOC>, </DOC>, <DOCNO>, and </DOCNO> markers.

Figure 2 provides an example of a filter for TREC Topic 173 which appears in Figure 1. The INF file in Figure 2 indicates the topic number, the size of the text window to be used, the number of synonym and domain files, an output filename, the actual file names of the synonym and domain files, a minimum document relevancy value to consider valid, and the number of combinations followed by the weighted combinations that could indicate a relevant piece of information.

Figure 2 also reveals the synonym and domain files which are specified in the INF file. There are five files specified and they are named to somewhat indicate their content. The filter is looking for the following information:

a. The name of an action.
b. The name of an affected area.
c. A synonym for the word "ban".
d. The name of an indicator for human use of tobacco.
e. A descriptive phrase for a smoking ban.

A central part of the scanner is the "text window". This structure is essentially an array which contains the current group of words being evaluated for local purposes. At any given point in processing, the text window contains the last *X* words read from the text, where *X* is specified as the window size in the INF file. For Topic 173, *X* is 300 as shown in Figure 2. This provides for a variety of local evaluation sizes (e.g., searching on a paragraph-by-paragraph size of roughly 100 words, as opposed to a sentence-by-sentence search of 20 words at a time). The text window's usage gives rise to the terms local and global. Local refers to an evaluation done exclusively on the text within the window, and global refers to an evaluation on the entire text of a document.

Documents are evaluated in a single-pass through the text. Document text begins immediately after the document ID, and ends at the document end marker. As each new word is scanned into the text window, it is compared to the entries in the files. This is accomplished by having read all the file entries into a memory resident hash table prior to the scanning process. If a match or matches are found, they are tallied in an array which contains the number of matches currently within the window, by file. At the same time, match counts that are passing out the end of the text window are subtracted from the array.

When the current word registers a match, there is an immediate evaluation of the current window's contents. For each valid combination specified in the INF file, there is determined a combination value. The combination value is the sum of the quotients of the number of non-zero, required matches (zero matches or non-required files add zero to the sum) for each file multiplied by 1 minus the result of 1 divided by the sum of the "parts" specified in the INF file for each particular file in the combination being evaluated plus the total number of files.

Only the highest combination value encountered is retained, such that at the end of the document, there will be a set of combination values which are maximums for the entire document. Concurrent with these local evaluations, a global document match array is maintained for combined local and global weighting at the end of the document.

Once the entire text of a document has been scanned, a local weight is determined by summing the squares of the best combination weights achieved within the document. Following this, a series of combination values are calculated, then summed, and squared to arrive at a global weight. This operation is identical to the combination calculations for local weight except that the global match count array is used instead of the array for the window. The final weight of the document is then reported as 75% of the local weight added to 25% of the global weight.

The filter shown in Figure 2 was built in two hours by Kai-Lin Hu. Several existing files from the filter for Topic 125 were used. Entries in the few files created for this filter are words and phrases (and their variants) found by reading some of the known relevant documents (the training data) for this topic. This filter achieved best performance for Topic 173.

Topic 173 - Smoking Bans

INF File For Topic 173

```
173
300
5
t173.out 0
t173.action_name.dom 1
t173.affected_area.dom 1
t173.ban.syn 1
t173.human_use.dom 1
t173.smoking_ban.dom 1
0.0
1
5 1 1 5 13
```

t173.action_name.dom
name of an action
(domain file - from Topic 125 filter)

```
CURTAILED ADVERTISING,
NON SMOKING AREA,
NON SMOKING AREAS,
NON-SMOKING AREA,
NON-SMOKING AREAS,
SPECIAL TAX,
SPECIAL TAXES,#
```

t173.afftected_area.dom
name of an affected area
(domain file - specified in Topic 173)

```
BANKS,
DEPARTMENT STORE,
DEPARTMENT STORES,
DOMESTIC FLIGHTS,
ELEVATOR,
ELEVATORS,
HOSPITAL,
HOSPITALS,
HOTEL,
HOTELS,
LARGE STORE,
LARGE STORES,
RESTAURANT,
RESTAURANTS,
TAXICAB,
TAXICABS,
THEATER,
THEATERS,
THEATRE,
THEATRES,
WORKPLACE,
WORKPLACES,#
```

t173.ban.syn
synonym for ban
(synonym file - specified in Topic 173)

```
BAN,
BANNED,
BANNING,
BANS,
FINE,
PENALIZE,
PENALIZED,
PENALIZES,
PENALIZING,
PENALTY,
PROHIBIT,
PROHIBITED,
PROHIBITING,
PROHIBITS,
RESTRICT,
RESTRICTED,
RESTRICTING,
RESTRICTION,
RESTRICTIONS,
RESTRICTS,#
```

t173.human_use.dom
name of indicator for human use of tobacco
(domain file - from Topic 125 Filter)

```
CIGARETTE,
CIGARETTES,
SMOKE,
SMOKED,
SMOKER,
SMOKERS,
SMOKES,
SMOKING,#
```

t173.smoking_ban.dom
descriptive phrase for a smoking ban
(domain file - specified in Topic 173)

```
AGAINST SMOKER,
AGAINST SMOKERS,
AGAINST SMOKING,
ANTI SMOKING,
ANTI-SMOKING,
ATTACK ON SMOKING,
ATTACK SMOKING,
ATTACKED SMOKING,
ATTACKING SMOKING,
ATTACKS ON SMOKING,
ATTACKS SMOKING,
BAN ADVERTISEMENT BY TOBACCO,
```

Figure 2. A Filter for TREC Topic 173 (continued on next page).

BAN ADVERTISEMENTS BY TOBACCO,
BAN CIGARETTE SMOKING,
BAN ON PUBLIC SMOKING,
BAN ON SMOKING,
BAN SALES OF CIGARETTES,
BAN SMOKING,
BAN THE SALE OF KENTS,
BANNED CIGARETTE SMOKING,
BANNED SMOKING,
BANNING CIGARETTE SMOKING,
BANNING SMOKING,
BANNING THE SALE OF KENTS,
BANS ADVERTISEMENT BY TOBACCO,
BANS ADVERTISEMENTS BY TOBACCO,
BANS CIGARETTE SMOKING,
BANS ON PUBLIC SMOKING,
BANS ON SMOKING,
BANS SMOKING,
CIGARETTE BAN,
CIGARETTE BANS,
CIGARETTE-BAN,
CIGARETTE-BANS,
CIGARETTES OUT OF THE WAY,
FORBID SMOKING,
FORBIDDEN SMOKING,
FORBIDDING SMOKING,
FORBIDS SMOKING,
INCREASE TAXES ON TOBACCO,
INCREASED TAXES ON TOBACCO,
INCREASES TAXES ON TOBACCO,
INCREASING TAXES ON TOBACCO,
NO SMOKING RULE,
NO SMOKING RULES,
NO-SMOKING DAY,
NO-SMOKING DAYS,
NO-SMOKING RULE,
NO-SMOKING RULES,
NON SMOKERS EQUAL RIGHT,
NON SMOKERS EQUAL RIGHTS,
NON-SMOKERS EQUAL RIGHT,
NON-SMOKERS EQUAL RIGHTS,
NON-SMOKING DAY,
NON-SMOKING DAYS,
NON-SMOKING MOVEMENT,
NON-SMOKING MOVEMENTS,
NONSMOKING DAY,
NONSMOKING DAYS,
NONSMOKING MOVEMENT,
NONSMOKING MOVEMENTS,
NOT TO SMOKE,
OFF LIMITS TO SMOKERS,
OFF-LIMITS TO SMOKERS,
ONLY FOR NON-SMOKER,
ONLY FOR NON-SMOKERS,
OUTLAW TOBACCO,
PENALIZE SMOKER,
PENALIZE SMOKERS,
PENALIZES SMOKERS,
PENALIZING SMOKERS,
PROHIBIT ADVERTISING OF CIGARETTES,
PROHIBIT SMOKING,
PROHIBIT TEEN-AGER SMOKING,
PROHIBIT TEEN-AGERS FROM SMOKING,
PROHIBIT TEENAGER SMOKING,
PROHIBIT TEENAGERS FROM SMOKING,
PROHIBIT TOBACCO,
PROHIBITED SMOKING,
PROHIBITED TOBACCO,
PROHIBITING SMOKING,
PROHIBITING TOBACCO,
PROHIBITION AGAINST CIGARETTE SMOKING,
PROHIBITION AGAINST SMOKING,
PROHIBITIONS AGAINST CIGARETTE
SMOKING,
PROHIBITIONS AGAINST SMOKING,
PROHIBITS ADVERTISING OF CIGARETTES,
PROHIBITS SMOKING,
PROHIBITS TOBACCO,
RAISE TAXES ON TOBACCO,
RAISING TAXES ON TOBACCO,
RESTRICT THE RIGHTS OF SMOKERS,
RESTRICTED THE RIGHTS OF SMOKERS,
RESTRICTING THE RIGHTS OF SMOKERS,
RESTRICTS THE RIGHTS OF SMOKERS,
SMOKE FREE ENVIRONMENT,
SMOKE FREE ENVIRONMENTS,
SMOKE FREE NORWAY,
SMOKE FREE,
SMOKE-FREE,
SMOKELESS,
SMOKING BAN,
SMOKING BANS,
SMOKING HAS BEEN BANNED,
SMOKING IS ALSO PROHIBITED,
SMOKING IS BANNED,
SMOKING IS PROHIBITED,
SMOKING RESTRICTION,
SMOKING RESTRICTIONS,
SMOKING WAS BANNED,
SMOKING WILL BE BANNED,
SMOKING-BAN,
SMOKING-BANS,
STASH THE CIGARETTE,
STASH THE CIGARETTES,
STOP SELLING TOBACCO,
STOP-SMOKING,
STOPPED SELLING TOBACCO,
STOPPING SELLING TOBACCO,
STOPS SELLING TOBACCO,#

Figure 2 (continued from previous page).

Spanish Adhoc Retrieval

Our participation in the Spanish adhoc task originated as Sara Abbott's semester project for an undergraduate course in *Data Processing Systems Implementation*. During the 1995 spring semester, our filtering system was modified to handle the special 8-bit characters found in the Spanish text, and an auxiliary program was written to "expand" Spanish verbs and adjectives listed in synonym and domain files in their infinitive and masculine singular forms, respectively. As a summer semester independent study project, Sara developed filters (synonym, domain, and INF files) for one run on Topics SP1-SP25, and two runs on Topics SP26-SP50. Sara speaks Spanish as a second language. She had previous experience submitting TREC results by participating in the TREC-3 Category B routing task using her own text scanning system; however, she had no previous experience using our present filtering system.

Modification to Accommodate Special 8-bit Characters

In the process of *ftp*-ing and decompressing the approximately 200 megabytes of Spanish text, the 8-bit characters relevant to our task, *á, é, í, ó, ú, ü, ñ*, and *Ñ*, were translated to other extended *ASCII* characters, such as Greek letters and mathematical symbols. Rather than investigate the cause of this undesired translation, we chose to simply translate the symbols to appropriate uppercase unaccented and unumlauted vowels and to restore the letter *Ñ* in its uppercase form. This works because our filtering system converts all of the regular ASCII characters to uppercase before making hash table comparisons.

We treat accented and umlauted vowels as regular vowels because accent marks and umlauts in Spanish generally serve no other purpose than to indicate an irregularly accented syllable, an irregular pronunciation, or a variation in verb form. These diacritical marks are also often inadvertently or deliberately omitted in Spanish text. We found this to be the case, at times, in the *El Norte* collection. It should be noted that this collection contains no uppercase versions of accented or umlauted vowels, possibly due to the unavailability of some of these characters in the keyboard configuration used to create the text. Since our system ignores most punctuation, we ignore the *¿* and *¡* symbols. Only six lines were added to the source code of our filtering system to handle Spanish special characters.

Auxiliary Program for Conjugation of Spanish Verbs

Perhaps the most laborious part of the adaptation of our filter system to accommodate Spanish was the creation of an auxiliary program to generate various Spanish verb forms, when given the infinitive form. This was implemented by allowing placement of the infinitive form, followed by an asterisk flag, in a preliminary synonym or domain list. Given the preliminary list as standard input, a *flex* program, which we call *lexpand*, produces an equivalent longer list containing all useful forms of "flagged" infinitives. Words and phrases which are not flagged remain unaltered. Flagged adjectives ending in *O* are expanded to include feminine and plural forms. Appendix A shows an example of a preliminary synonym list, along with a verb and an adjective from the list as expanded by *lexpand*.

Since *lexpand* at times generates nonsensical verb forms which would never occur (*comermelo*, for example), and includes *vosotros* forms, which are not generally used in Mexico, other than in Biblical references, our philosophy is essentially "better too much than not enough." The junk verbs that we generate only slightly hinder efficiency. The advantage of this approach over stemming is that we can edit the lists generated by *lexpand* to exclude any verb that we do not like. In this manner we avoid some of the following incorrect identifications:

Problem Word	English Meaning	Mistaken For	English Meaning
para	for	*para*	stops
cómo, como	how, like	*como*	eat
nada	nothing	*nada*	swims
vino	wine	*vinó*	came

The *lexpand* program, in its present state, consists of about 700 lines of *flex* source code, which, in turn, generates about 2500 lines of *C* code. It could be improved if broken up into more modules, and through the use of more tables. *Flex* was chosen over other available lexical analyzers because it generates fast executable programs and has fewer restrictions on the size of the tables that it produces, since it dynamically allocates table space.

Construction of Spanish Queries

Spanish queries were constructed manually, and like our English queries (filters) for the routing task, consisted of collections of key phrases and words, referenced by INF files. Although no actual documents were read as we formed TREC-4 Spanish queries, some phrase collections were developed by extracting single lines of text containing key words from the Spanish adhoc document collection and then building lists of phrases with similar meanings. Synonym and domain lists were developed with the help of friends in the local Latin-American community, and with the aid of the following dictionaries and thesauri:

Diccionario de Sinónimos Explicados,
Alonzo, Martín [1].

Larousse Diccionario Escolar,
García-Pelayo y Gross, Ramón [7].

The New World SPANISH-ENGLISH and ENGLISH-SPANISH Dictionary,
Ramondino, Salvatore (editor) [10].

We found an abundance of information related to Mexican commerce and politics, and even some specific names of corporations and trade promotion organizations in the following reference:

Mexico Business: The Portable Encyclopedia for Doing Business with Mexico,
Nolan, James L. etal [9].

Description of Spanish Adhoc Runs

Our two runs for Spanish Topics 26-50, UCFSP1 and UCFSP2, use essentially the same collections of phrases and words for each query. UCFSP1, which was our primary run for official evaluation, is a conservative run, using only single weighting patterns and no negative weighting. UCFSP2 was made using a slight modification to the filter operation described in an earlier section. The modification was to select the highest pattern weight for each window as that window's weight, and to use the highest global weight generated by any pattern. Seventy-five percent of the square of the window weight was then added to twenty-five percent of the square of the global weight to yield a total weight for the given document.

The filter operation used for the other Spanish runs (and the English runs) takes the sum of the squares of all pattern weights for both the window weight and global weight, rather than choosing the highest weights. Squaring the highest pattern weight instead of summing the squares of all pattern weights means that the scanner is now choosing between patterns rather than "averaging" all patterns within a given document.

The filter operation for the UCFSP2 run also allows for lists to receive negative weights within patterns. The assignment of negative weight to a list seeks to increase precision, by subtracting weight when certain words or phrases are encountered. The strategy is that the presence of certain words or phrases can hint that positively weighted words or phrases are being taken out of context. Topic SP41, for example, deals with measures taken in Mexico to control or limit flooding. Since there was extensive flooding along the Mississippi River during 1993, it would be reasonable to expect a query for words equivalent to *flood* and *flooding* to retrieve some documents pertaining to flooding in the United States. A list was constructed for this topic consisting of the names of states along the Mississippi and the list was assigned a negative weight.

Problemas, Problemas, y Más Problemas

As soon as we had our filter system Spanish-ready, we began making filters for the Spanish Topics SP1-SP25 from TREC-3 because qrels were available for these topics. Chris Buckley from Cornell sent us the evaluations from one of their runs to use as a benchmark. Our performance was, generally, considerably lower. However, we found that by eliminating entire domain and synonym lists in a query (filter), we could match or exceed the benchmark. On one topic, a precision of above .8900 was achieved using a single synonym list containing just six words! To us, this meant that the qrels were not very useful.

Even so, we thought that we could use the.qrels to at least get some indication of our performance. This was not the case. After reading only a few of the not-relevant documents that we retrieved, and relevant documents that we did not retrieve, it became obvious that running the evaluator using judgments from TREC-3 gave virtually no real indication of performance. We were actually making our queries worse as we tried to approach benchmark precision.

We had expected to retrieve some documents that were not retrieved last year, due to the small number of participants last year. What we didn't expect were the many not-relevant documents that were judged to be relevant in the documents that we examined. Topic SP5 deals with *maquiladoras*, or in-bond manufacturing enterprises. It is understandable how some scanning methods, particularly those using trigrams, might pick up a document containing the word *maquillaje*, meaning "the application of cosmetics," but it was not expected that such a document could be judged to be relevant. Document SP94-0032014, which was judged to be relevant for Topic SP5, is entitled "*Ofrece clases de maquillaje*," or "Cosmetology classes are offered," and goes on to describe the classes offered.

This is perhaps an extreme example of how odd the TREC-3 judgments were. All of the topics that we looked at had many not-relevant documents judged to be relevant. Usually the lack of relevancy was a bit more subtle. Topic SP1, for example, dealt with Mexican opposition to NAFTA. We found that there were more documents dealing with U.S. or Canadian opposition, or that were neutral or pro-NAFTA than there were truly relevant documents in the ones rated as relevant. It appears that if the document contained a reference to NAFTA, it was judged as relevant.

Unfortunately, we did not have time to do a really good set of runs for the Topics SP1-SP25, after the discovery that the qrels were not useful. The run which we submitted, UCFSP0, was made with very few, very short synonym and domain lists, but should still be somewhat useful in the development of new qrels.

Another problem which we encountered with two topics, one from Topics SP1-SP25, and the other from Topics SP26-SP50, was that the event which the topic dealt with occurred after the documents in the collection were written! The *El Norte* collection appears to be dated from December of 1992 through September of 1993. Topic SP2 dealt with the political effects of the assassination of Mexican presidential candidate Luís Donaldo Colosio Murrieta, who was slain on March 23, 1994 [6]. There can be no relevant documents on this topic, although one was found last year. Topic SP31 dealt with measures taken by the Mexican government to resolve a "dispute" in the Mexican state of Chiapas, which actually began on New Year's Day of 1994, when the Zapatista National Liberation Army declared war against the government and seized four towns [6]. There are documents in the collection which mention rebels in Chiapas, and that are related to Topic SP13, which deals with confrontations between the Mexican army and suspected Zapatista rebels, but it is our belief that the "dispute" referred to in the topic description was actually intended to mean the full scale dispute that erupted on January first.

As evidenced above, by the questionable meaning of the word "dispute," the shortened topic descriptions for TREC-4 Spanish topics are somewhat vague, and at times we were unsure what to look for. The description for Topic SP42, as another example, consists of only the question, "Will NAFTA be successful in Mexico?" We didn't know whether to look for documents stating that NAFTA would be a success, or those just stating an opinion one way or the other. We felt that the descriptions for Topics SP1-SP25 were a lot more specific and easier to work with than those for Topics SP26-SP50.

Network Environment

The computer network that enabled undergraduate students to develop filters for TREC-4 is the same one we used for TREC-3 [4].

For training routing filters on just the Vol. 1 and Vol. 2 CDs:

1. The Vol. 1 CD was copied to the hard drive of a PC running Linux (a public domain version of Unix) and functioning as an NFS node on the network.

2. The Vol. 2 CD was copied to the hard drive of a SPARC Server 690MP (four processors) on the network.

3. Students ran filters and viewed training text from 32 RISC machines across the network.

For the UCF100, UCFSP0, UCFSP1, and UCFSP2 runs:

1. The routing and Spanish adhoc document collections were copied to the hard drive of the SPARC Server 690MP (four processors) on the network.

2. Most filters were run on the SPARC Server 690MP. (A few were run on the RISC machines.)

Each of the 32 RISC 6000 machines had 16 MB of RAM. The NFS node had 16 MB of RAM, and the SPARC Server 690MP had 128 MB RAM. All these machines (except for the NFS node) were shared with normal University and Department computing.

During training in the spring semester, students began to submit multiple filters. This was OK as long as the filters ran on different RISC machines. But some students did not pay attention and they started submitting multiple filters on a RISC machine. Many of the filters were doomed because they were not set up properly. This caused severe network problems and it was difficult to even try to log into some of the RISC machines on many occasions. Other students started rebooting the RISC machines and this stopped many filter runs.

So, during the summer semester, only Dr. Driscoll could activate a filter run on the RISC machines. Students had to meet Dr. Driscoll and he had to approve their filter files before he would run them. This solved all the network and machine load problems, but required that Dr. Driscoll be available about 8 hours a day just to look at and run filters.

Performance Results and Analysis

For the Category A routing task, fifty filters were developed for one routing run (UCF100). A large amount of time went into the development of 28 filters. The other 22 filters were developed rather quickly in the last few weeks before we turned in our routing queries. For just one of the topics, our filtering approach had the best average precision. For 16 of the topics, our average precision was above the median; for 34 of the topics, our average precision was below the median. Our overall average precision was .2285 for this experiment.

This performance is a lot worse than our performance for the routing experiment last year. We believe the following contributed to this year's lower performance:

1. We did not use training data from the Vol. 3 CD.

2. The Ziff document collection on the Vol. 3 CD was part of the "new" routing document collection.

3. We believed a lot of the training data was incorrect and we made our filters retrieve what we believed to be relevant, not what the assessor considered relevant.

Not using the training data on the Vol. 3 CD was a mistake because it is probably the most accurate of the training data. We should have used it. But we believe the real problem is that combined with the fact that the Ziff document collection on the Vol. 3 CD was also part of the "new" routing document collection! We developed our filters from March to late July and sent in our routing queries during the last week of July. That is when we discovered that the Ziff document collection on the Vol. 3 CD was part of the "new" routing document collection. Topics in the range of Topic 051 through Topic 150 had training data on the Vol. 3 CD. For all but one of the routing topics chosen this year in that range, our filters performed below the median. The highest performance was for our Topic 117 filter which performed right at the median this year. We used our Topic 117 routing filter from last year without making a change to it; last year it had the best performance! So, we believe item 1 and item 2 above really hurt our performance. We also believe item 3 above was a factor causing our lower performance.

For the Spanish adhoc task, two runs were made. The UCFSP1 run was our primary run to be evaluated. The difference between the UCFSP2 run and the UCFSP1 run is explained in an earlier section. The performance of the UCFSP1 run is extremely good. It had the best average precision for 17 of the 25 topics. On all but one topic, Topic SP45, average precision was above the median. The overall average precision was .4918 for this experiment. The overall average precision of the UCFSP2 run was just slightly lower than that of UCFSP1; and, for most topics, the UCFSP2 performance was just below the UCFSP1 performance; however, UCFSP2 also had the best average precision for 3 more of the 25 topics. So, our filtering approach generated best average precision performance for 20 of the 25 Spanish adhoc topics.

Negative weights were used in fourteen of the twenty-five UCFSP2 filters, and generally lowered precision, when compared to corresponding UCFSP1 runs in which no negative weighting was used. For two topics, Topic SP37 and Topic SP41, higher precision was obtained using negative weights. The logic behind the negation used in Topic SP41 was explained earlier, and the negation used in Topic SP37 will be described in a later section, where Topic SP37 was chosen for a multilingual experiment. These mixed results suggest that in an adhoc environment, where no training is permitted, negation must be used very carefully, if at all.

Multiple patterns, used in seven of the UCFSP2 filters, were also generally unsuccessful. A higher precision was attained by the use of multiple patterns for only Topic SP50. In that particular case, the weighting of the corresponding UCFSP1 filter was far from optimum. The topic dealt with "The Fabrication of Gold and Silver Jewelry in Mexico." Higher weight should have been assigned to the synonym list for the entity "jewelry" than to the domain list associated with the jewelry attribute "type of metal." We later obtained a precision of above .8 for Topic SP50 by changing the weights used in the UCFSP1 filter submitted for Topic SP50. Our official precision for this topic was .3358 for UCFSP1 and .4750 for UCFSP2, both above the median, but somewhat lower than that of the best performing filter. It appears that poor weighting was a factor in most cases where our performance was not so good.

A Multilingual Experiment

After submitting our Spanish runs, we began reading some of the documents that we had retrieved as relevant to get some indication of our Spanish Adhoc performance. We noticed that there were many documents in the *El Norte* collection that were relevant to English topics. Since the filtering system had already been modified to accept multiple patterns, the idea occurred to us that possibly a single filter could be used to scan both the English and Spanish collections in one single run, producing a single ranked list.

The idea was extended to include test documents in French and German when an article relevant to Topic SP37 (Evidence of Aztec Heritage and Culture in Mexico) was found in the magazine *Quinto Lingo* [8]. The article had been translated into English, Spanish, French, and German versions. All four versions of the article, "Who was Quetzalcoatl?" were typed in <SGML> format as separate documents, and put into a file to be scanned along with the Volume 1 CD, the Volume 2 CD, and the *El Norte* collection. Accented and umlauted characters were typed as regular unaccented characters. It should be noted that no additional French or German text was scanned, mainly due to lack of time and a desire not to infringe in any manner on copyrights.

Since the UCFSP2 query for Topic SP37, which uses negative weights, had the best precision and recall for that topic, it was decided that the closest translations of that query that were possible, given our limited familiarity with French and German, would be used to create separate synonym and domain lists for each additional language. Small allowances were made for differences in verb conjugation, noun and adjective declension, gender, and the way in which possessives are formed. The German lists reflect a non-speaker's humble attempt to master the grammar of a relatively complex language in a few hours. Declension of German nouns and adjectives is incomplete. Possessives using *of*, and equivalent prepositions in other languages (*de*, *du*, *auf*, etc.) were not included, as that form of possessive was overlooked in the original Spanish query. Accents and umlauts were removed. See Appendices B-E for the actual synonym and domain lists that were used for the various languages.

A single INF file was created called "multi.inf." The version presented at NIST was the version we actually ran, which used DOS compatible file names. Here we present an easier to understand version:

```
37                            (topic number)
200                           (window size)
multi.out                     (the ranked list)
Aztec_culture.syn.Span        (Spanish synonym files)
Aztec.syn.Span
culture.syn.Span
Aztec_culture.syn.Eng         (English synonym files)
Aztec.syn.Eng
culture.syn.Eng
Aztec_culture.syn.Fren        (French synonym files)
Aztec.syn.Fren
culture.syn.Fren
Aztec_culture.syn.Ger         (German synonym files)
Aztec.syn.Ger
culture.syn.Ger
evidence.dom.Nah              (domain file in the Aztec language, Nahuatl)
garbage.dom.Span              (a list of Spanish words we don't want)
0.00                          (lowest weighting to be output)
4                             (number of patterns)
6 3 2 0 0 0 0 0 0 0 0 0 1 -3  (pattern for Spanish)
0 0 0 6 3 2 0 0 0 0 0 0 1 -3  (pattern for English)
0 0 0 0 0 0 6 3 2 0 0 0 1 -3  (pattern for French)
0 0 0 0 0 0 0 0 0 6 3 2 1 -3  (pattern for German)
```

As can be seen from the patterns, equivalent synonym lists are given the same weight for each language.

The Nahuatl words in "evidence.dom.Nah" are, for the most part, the names of Aztec, Toltec, and Mayan deities. They were put in a domain file because "deity" could be looked upon as a subset of a "type" attribute for the entity "evidence." "evidence.dom.Nah" was given weight in the patterns for all four languages because we presumed that there would be no attempt to convert these names to another language. We did expect some misspellings and we found some when reading the documents that we retrieved.

The "garbage.dom.Span" file is the same as used in the original Spanish query. It seeks to avoid documents pertaining to "Aztec Stadium," and "Aztec Avenue" (in Monterrey, Nuevo Leon, where *El Norte* is published). After reading some retrieved documents, we discovered that there is also an "Aztec Television" in Mexico, which we could have included in this negatively weighted list. We did not attempt to translate this list to other languages, but gave it the same negative weight in all patterns (if a phrase from this list showed up in a non-Spanish document it would still warrant a negative weight). We could have chosen to give it a zero weight in non-Spanish patterns.

We ran our multilingual filter across Vol. 1 and Vol. 2 CDs of the English collection, the *El Norte* collection, and the test file from *Quinto Lingo* in about six hours. The top 25 documents are as follow:

37 Q0 SP94-0000465	0	0.87533 UCFSP2	
37 Q0 WSJ910417-0120	1	0.86645 UCFSP2	
37 Q0 FRA000000-0001	2	0.86320 UCFSP2	(French document from *Quinto Lingo*)
37 Q0 SP94-0037330	3	0.86112 UCFSP2	
37 Q0 SP94-0043006	4	0.86048 UCFSP2	
37 Q0 SP94-0110606	5	0.86011 UCFSP2	
37 Q0 SP94-0005842	6	0.73133 UCFSP2	
37 Q0 AP881027-0004	7	0.72425 UCFSP2	
37 Q0 AP881020-0253	8	0.72425 UCFSP2	
37 Q0 SP94-0202852	9	0.72312 UCFSP2	
37 Q0 SP94-0000857	10	0.72094 UCFSP2	
37 Q0 SP94-0110560	11	0.72020 UCFSP2	
37 Q0 AP890828-0214	12	0.72008 UCFSP2	
37 Q0 AP890728-0053	13	0.71979 UCFSP2	
37 Q0 SP94-0101980	14	0.71615 UCFSP2	
37 Q0 ENG000000-0001	15	0.21872 UCFSP2	(English document from *Quinto Lingo*)
37 Q0 DEU000000-0001	16	0.21823 UCFSP2	(German document from *Quinto Lingo*)
37 Q0 ESP000000-0001	17	0.21790 UCFSP2	(Spanish document from *Quinto Lingo*)
37 Q0 SP94-0033589	18	0.21758 UCFSP2	
37 Q0 SP94-0110291	19	0.21748 UCFSP2	
37 Q0 SP94-0107105	20	0.21740 UCFSP2	
37 Q0 AP880314-0254	21	0.21706 UCFSP2	
37 Q0 SP94-0038957	22	0.21700 UCFSP2	
37 Q0 SP94-0003010	23	0.21669 UCFSP2	
37 Q0 WSJ910905-0078	24	0.21609 UCFSP2	

As seen above, the four known relevant documents from *Quinto Lingo* all appeared in our top 25 documents. The French document received a higher rank than its counterparts due to a difference in the way a possessive form was expressed in the translation. The top 25 also included 14 Spanish documents, 13 of which were judged to be relevant, and 7 English documents, 6 of which we considered relevant. Of the 1000 documents that we ranked, 111 were from the Vol. 1 and Vol. 2 CDs. Although we did not take the time to read all of these documents, it is interesting to note that we had document identifiers from every directory on the Vol. 1 and Vol. 2 CDs in our ranked top 1000.

This experiment sidesteps some important aspects of the grammars involved, particularly verb conjugation, and lacks an adequate test collection of French and German documents. It does, however, serve to illustrate that our filtering system is very flexible. To accurately scan text in French or German it would be necessary to modify the filter somewhat to deal with special French accents, and an additional German character resembling *β*. Auxiliary programs could be built to conjugate at least the regular verbs of both languages, and possibly to deal with some of the noun and adjective declension seen in German. Some care would have to be taken to avoid using words in lists for one language that would mean something else in another, or some method would have to be devised to automatically determine the language of a document, and which pattern to use. We feel that Italian and Portuguese could also be filtered with only a small amount of alteration to our system.

Acknowledgments

Sara Abbott performed all of the Spanish experiments, as well as the multilingual experiment. Kai-Lin Hu created 22 filters during the few weeks just before routing queries were submitted to NIST. In some awful confusion in August, some of her very good filters were accidentally not used for submission of final results.

Michael Miller studied the source code of our filter system and made himself available for questions about its operation. He also developed useful utility programs for evaluating filter runs and retrieving documents. Gary Theis is the original author of the filter system source code. For the routing experiments, the relevancy evaluation within the filter system was not changed from that developed by Gary.

The following students worked very hard developing filters for the routing topics: Fernando DaSilva, Richard Edmundson, Fritz Feuerbacher, Carl Free, Keith Hamelin, Brian Hinken, Hong Ji, John Kuhn, Stewart Ort, Gladys Otegbeye, Xiang Gen Qian, Ed Saab, Mark Tootle, and Ahmed Wreiden. Alicia Chea helped with the French synonym and domain files used in the multilingual experiment.

Chris Buckley provided the evaluation of one of Cornell's TREC-3 Spanish adhoc runs so that we would have a benchmark for initial experiments involving Topics SP1-SP25.

También queremos agradecer a nuestros amigos de la comunidad latina quienes han ofrecido sus consejos y nos han ayudado en varias otras maneras. Se incluyen a Bernardo Arteaga, José y Ilda Ayala, Luís Castañeda, Miguel de Jesús Ruíz Reyes, y Salvador Ruíz Reyes, todos del lindo país de México, asi como a Roger Huamán y Peggie Castle Haas, de Lima Perú. Sobre todo agradecemos a la Dra. Flor de María Gallardo Vásquez, también de Lima, por haber leído tantos documentos, dando su opinión sobre la pertinencia de todos. El éxito que tengamos con la colección "El Norte" lo debemos a estos queridos amigos.

References

[1] Martín Alonzo, *Diccionario de Sinónimos Explicados*, Madrid: EDAF, 1984.

[2] N. J. Belkin and W. Bruce Croft, "Information Filtering and Information Retrieval: Two Sides to the Same Coin?", *Communications of the ACM*, Vol. 35, No. 12, pp. 29-38, 1992.

[3] C. J. Date, *An Introduction to Database Systems (Sixth Edition)*, Addison-Wesley Publishing Company, Inc., 1995.

[4] J. Driscoll, G. Theis and G. Billings, "Using Database Schemas to Detect Relevant Information", *Proceedings of the Third Text Retrieval Conference (TREC-3)*, NIST Special Publication 500-225 (D. K. Harman, ed.), April 1995.

[5] R. Elmasri and S. B. Navathe, *Fundamentals of Database Systems (Second Edition)*, The Benjamin/Cummings Publishing Company, Inc., 1994.

[6] Robert Famighetti, ed., *The World Almanac and Book of Facts*, Funk & Wagnalls, pp.46-53, 1994.

[7] Ramón García-Pelayo y Gross, *Larousse Diccionario Escolar (First Edition)*, México 06600, D.F., Larousse, 1987.

[8] Lowell Harmer, "Who Was Quetzalcoatl?", *Quinto Lingo, The Multi-lingual Magazine*, Vol. 15, No. 1, American National Heritage Assoc., Arlington VA, September 1977.

[9] James L. Nolan etal., *Mexico Business: The Portable Encyclopedia for Doing Business with Mexico*, World Trade Press, San Rafael, CA, 1994.

[10] Salvatore Ramondino, ed., *The New World SPANISH-ENGLISH and ENGLISH-SPANISH Dictionary*, Signet, New York, 1991.

APPENDIX A

EXPANSION OF VERBS AND ADJECTIVES

A sample preliminary synonym list

ANALISIS,
ANALIZAR*,
CLARIFICAR*,
CLARIFICACION,
CONCLUIR*,
CONCLUSION,
CONCLUSIONES,
CURAR*,
CURACION,
CURACIONES,
DESARROLLAR*,
PROCESAR*,
DESCUBRIR*,
HALLAR*,
EVIDENCIAR*,
MOSTRAR*,
MOSTRACION,
DEMONSTRAR*,
DEMOSTRACION,
PROBAR*,
PROBACION,
EXPERIMENTAR*,
EXPERIMENTACION,
INTRODUCIR*,
INTRODUCCION,
LLEVAR*,
CONDUCIR*,
CONDUCCION,
INVESTIGAR*,
INVESTIGACION,
RECONOCER*,
ESCRUDIÑAR*,
INDAEGAR*,
SONDEAR*,
REVOLUCIONAR*,
REVOLUCIONACION,
ILUMINAR*,
ILUMINACION,
ALUMBRAR*,
ENFOCAR*,
ESTUDIAR*,
ESTUDIOS,
TECNICO*,
TECNOLOGO*,
TECNOLOGIA,
TECHNOLOGIAS,
EXAMINAR*,
EXAMEN,
EXAMENES,
EXAMINACION,
EXAMINACIONES,
TERAPIA,
TERAPIAS,
TRATAR*,
TRATAMIENTO,
TRATAMIENTOS,
COMPRENDER*,
ENTENDER*,
#

(This list was developed to test the *lexpand* program on English Topic 135.)

The verb *ANALIZAR as expanded**

ANALIZO,
ANALIZAS,
ANALIZA,
ANALIZAMOS,
ANALIZAIS,
ANALIZAN,

ANALIZABA,
ANALIZABAS,
ANALIZABAMOS,
ANALIZABAIS,
ANALIZABAN,

ANALICE,
ANALIZASTE,
ANALIZO,
ANALIZASTEIS,
ANALIZARON,

ANALIZARE,
ANALIZARAS,
ANALIZARA,
ANALIZAREMOS,
ANALIZAREIS,
ANALIZARAN,

ANALIZARIA,
ANALIZARIAS,
ANALIZARIAMOS,
ANALIZARIAIS,
ANALIZARIAN,

ANALICE,
ANALICES,
ANALICEMOS,
ANALICEIS,
ANALICEN,

ANALIZARA,
ANALIZARAS,
ANALIZARAMOS,
ANALIZARAIS,
ANALIZARAN,

ANALIZASE,
ANALIZASES,
ANALIZASEMOS,
ANALIZASEIS,
ANALIZASEN,

ANALIZAR,
ANALIZARSE,
ANALIZARSELO,
ANALIZARSELOS,
ANALIZARSELA,
ANALIZARSELAS,
ANALIZARME,
ANALIZARMELO,
ANALIZARMELOS,
ANALIZARMELA,
ANALIZARMELAS,
ANALIZARLO,
ANALIZARLOS,
ANALIZARLA,
ANALIZARLAS,
ANALIZARLE,
ANALIZARLES,
ANALIZARNOS,

ANALIZANDO,
ANALIZANDOSE,
ANALIZANDOSELO,
ANALIZANDOSELOS,
ANALIZANDOSELA,
ANALIZANDOSELAS,
ANALIZANDOME,
ANALIZANDOMELO,
ANALIZANDOMELOS,
ANALIZANDOMELA,
ANALIZANDOMELAS,
ANALIZANDOLO,
ANALIZANDOLOS,
ANALIZANDOLA,
ANALIZANDOLAS,
ANALIZANDOLE,
ANALIZANDOLES,
ANALIZANDONOS,

ANALIZADO,
ANALIZADOS,
ANALIZADA,
ANALIZADAS,

The adjective *TECNICO as expanded**

TECNICO,
TECNICOS,
TECNICA,
TECNICAS,

APPENDIX B

SYNONYM AND DOMAIN FILES REFERENCED BY SPANISH PATTERN OF MULTILINGUAL EXPERIMENT

(weights indicated beside each file name)

<evidence.dom.Nah> 1

NETZAHUALCOYOTL,
MOTECUHZOMA,
XOCOYOTZIN,
TLALOC,
NAHUATL,
TEMPLO MAYOR,
MOCTEZUMA,
XIPE,
CAMAXTLI,
CAMAXTLI-MIXCOATL,
MICTLAN,
MICTLANTECUHTLI,
MICTLANCIUATL,
COATLIQUE,
COYOLXAUHQUI,
HUIZILOPOCHTLI,
HUITZILOPOCHTLI,
XIPETOTEC,
MIXCOATL,
TEZCATLIPOCA,
TENOCHTITLAN,
OMETECHTLI,
XOCHIQUETZAL,
XOCHIPILLI,
XOCHIPILLI-CINTEOTL,
CINTEOTL,
MITOTE,
MITOTES,
MONTECZUMA,
MONTEZUMA,
QUETZALCOATL,
#

<garbage.dom.Span> -3

AVENIDA AZTECA,
AVE AZTECA,
AV AZTECA,
ESTADIO AZTECA,
#

<Aztec.syn.Span> 3

AZTECA,
AZTECAS,
AZTEQUISMO,
AZTEQUISMOS,
#

<culture.syn.Span> 2

CULTURA,
CULTURAS,
HERENCIA,
HERENCIAS,
CULTURAL,
CULTURALES,
HISTORIA,
HISTORIAS,
MITO,
MITOS,
MUSEO,
MUSEOS,
MITOLOGIA,
CODICE,
CODICES,
MONOLITO,
MONOLITOS,
TEMPLO,
TEMPLOS,
#

<Aztec_culture.syn.Span> 6

HERENCIA AZTECA,
CULTURA AZTECA,
INFLUENCIA AZTECA,
INFLUENCIAS AZTECA,
INFLUENCIAS AZTECAS,
INFLUYO AZTECA,
INFLUYOS AZTECA,
INFLUYOS AZTECAS,
HISTORIAS AZTECA,
HISTORIA AZTECA,
HISTORIAS AZTECAS,
TEMPLO AZTECA,
TEMPLOS AZTECA,
TEMPLOS AZTECAS,
MITO AZTECA,
MITOS AZTECA,
MITOS AZTECAS,
MITOLOGIA AZTECA,
MITOLOGIAS AZTECA,
MITOLOGIAS AZTECAS,
CODICE AZTECA,
CODICES AZTECAS,
CODICES AZTECA,
DEIDAD AZTECA,
DEIDADES AZTECA,
DEIDADES AZTECAS,
DIOS AZTECA,
DIOSES AZTECA,
DIOSES AZTECAS,
DIOSA AZTECA,
DIOSAS AZTECA,
DIOSAS AZTECAS,
#

APPENDIX C

SYNONYM AND DOMAIN FILES REFERENCED BY ENGLISH PATTERN OF MULTILINGUAL EXPERIMENT

(weights indicated beside each file name)

<evidence.dom.Nah> 1

NETZAHUALCOYOTL,
MOTECUHZOMA,
XOCOYOTZIN,
TLALOC,
NAHUATL,
TEMPLO MAYOR,
MOCTEZUMA,
XIPE,
CAMAXTLI,
CAMAXTLI-MIXCOATL,
MICTLAN,
MICTLANTECUHTLI,
MICTLANCIUATL,
COATLIQUE,
COYOLXAUHQUI,
HUIZILOPOCHTLI,
HUITZILOPOCHTLI,
XIPETOTEC,
MIXCOATL,
TEZCATLIPOCA,
TENOCHTITLAN,
OMETECHTLI,
XOCHIQUETZAL,
XOCHIPILLI,
XOCHIPILLI-CINTEOTL,
CINTEOTL,
MITOTE,
MITOTES,
MONTECZUMA,
MONTEZUMA,
QUETZALCOATL,
#

<garbage.dom.Span> -3

AVENIDA AZTECA,
AVE AZTECA,
AV AZTECA,
ESTADIO AZTECA,
#

<Aztec.syn.Eng> 3

AZTEC,
AZTECS,
#

<culture.syn.Eng> 2

CULTURE,
CULTURES,
HERITAGE,
HERITAGES,
CULTURAL,
HISTORY,
HISTORIC,
HISTORICAL,
LEGEND,
LEGENDS,
LEGENDARY,
MYTH,
MYTHS,
MYTHICAL,
MUSEUM,
MUSEUMS,
MYTHOLOGY,
MYTHOLOGIES,
CODEX,
CODICES,
MONOLITH,
MONOLITHS,
TEMPLE,
TEMPLES,
#

<Aztec_culture.syn.Eng> 6

AZTEC HERITAGE,
AZTEC CULTURE,
AZTEC INFLUENCE,
AZTEC INFLUENCES,
AZTEC HISTORY,
AZTEC LEGEND,
AZTEC LEGENDS,
AZTEC TEMPLE,
AZTEC TEMPLES,
AZTEC MYTH,
AZTEC MYTHS,
AZTEC MYTHOLOGY,
AZTEC MYTHOLOGIES,
AZTEC CODEX,
AZTEC CODICES,
AZTEC DEITY,
AZTEC DEITIES,
AZTEC GOD,
AZTEC GODS,
AZTEC GODDESS,
AZTEC GODDESSES,
#

APPENDIX D

SYNONYM AND DOMAIN FILES REFERENCED BY FRENCH PATTERN OF MULTILINGUAL EXPERIMENT

(weights indicated beside each file name)

<evidence.dom.Nah> 1

NETZAHUALCOYOTL,
MOTECUHZOMA,
XOCOYOTZIN,
TLALOC,
NAHUATL,
TEMPLO MAYOR,
MOCTEZUMA,
XIPE,
CAMAXTLI,
CAMAXTLI-MIXCOATL,
MICTLAN,
MICTLANTECUHTLI,
MICTLANCIUATL,
COATLIQUE,
COYOLXAUHQUI,
HUIZILOPOCHTLI,
HUITZILOPOCHTLI,
XIPETOTEC,
MIXCOATL,
TEZCATLIPOCA,
TENOCHTITLAN,
OMETECHTLI,
XOCHIQUETZAL,
XOCHIPILLI,
XOCHIPILLI-CINTEOTL,
CINTEOTL,
MITOTE,
MITOTES,
MONTECZUMA,
MONTEZUMA,
QUETZALCOATL,
#

<garbage.dom.Span> -3

AVENIDA AZTECA,
AVE AZTECA,
AV AZTECA,
ESTADIO AZTECA,
#

<Aztec.syn.Fren> 3

AZTEQUE,
AZTEQUES,
AZTECS,
AZTEC,
#

<culture.syn.Fren> 2

CULTURE,
CULTURES,
HERITAGE,
HERITAGES,
CULTUREL,
CULTURELS,
CULTURELLE,
CULTURELLES,
HISTOIRE,
HISTOIRES,
HISTORIQUE,
HISTORIQUES,
LEGENDE,
LEGENDES,
LEGENDAIRE,
LEGENDAIRES,
MYTHE,
MYTHES,
MYTHIQUE,
MYTHIQUES,
MUSEE,
MUSEES,
MYTHOLOGIE,
MYTHOLOGIES,
CODICE,
CODICES,
MONOLITHE,
MONOLITHES,
TEMPLE,
TEMPLES,
#

<Aztec_culture.Fren> 6

HERITAGE AZTEQUE,
CULTURE AZTEQUE,
INFLUENCE AZTEQUE,
INFLUENCES AZTEQUES,
HISTOIRE AZTEQUE,
HISTOIRES AZTEQUES,
LEGENDE AZTEQUE,
LEGENDES AZTEQUES,
TEMPLE AZTEQUE,
TEMPLES AZTEQUES,
MYTHE AZTEQUE,
MYTHES AZTEQUES,
MYTHOLOGIE AZTEQUE,
MYTHOLOGIES AZTEQUES,
CODICE AZTEQUE,
CODICES AZTEQUES,
DIEU AZTEQUE,
DIEUX AZTEQUES,
DEESSE AZTEQUE,
DEESSES AZTEQUES,
#

APPENDIX E

SYNONYM AND DOMAIN FILES REFERENCED BY GERMAN PATTERN OF MULTILINGUAL EXPERIMENT

(weights indicated beside each file name)

<evidence.dom.Nah> 1

NETZAHUALCOYOTL,
MOTECUHZOMA,
XOCOYOTZIN,
TLALOC,
NAHUATL,
TEMPLO MAYOR,
MOCTEZUMA,
XIPE,
CAMAXTLI,
CAMAXTLI-MIXCOATL,
MICTLAN,
MICTLANTECUHTLI,
MICTLANCIUATL,
COATLIQUE,
COYOLXAUHQUI,
HUIZILOPOCHTLI,
HUITZILOPOCHTLI,
XIPETOTEC,
MIXCOATL,
TEZCATLIPOCA,
TENOCHTITLAN,
OMETECHTLI,
XOCHIQUETZAL,
XOCHIPILLI,
XOCHIPILLI-CINTEOTL,
CINTEOTL,
MITOTE,
MITOTES,
MONTECZUMA,
MONTEZUMA,
QUETZALCOATL,
#

<garbage.dom.Span> -3

AVENIDA AZTECA,
AVE AZTECA,
AV AZTECA,
ESTADIO AZTECA,
#

<Aztec.syn.Ger> 3

AZTEK,
AZTEKEN,
AZTEKER,
AZTEKERN,
AZTEKISCH,
AZTEKISCHE,
AZTEKISCHEN,
AZTEKISCHER,
AZTEKISCHERN,
#

<culture.syn.Ger> 2

KULTUR,
KULTUREN,
BILDUNG,
BILDUNGEN,
ERBSCHAFT,
ERBSCHAFTEN,
KULTURELL,
KULTURZENTRUM,
KULTURZENTREN,
GESCHICHTE,
GESCHICHTEN,
GESCHICHTSWISSENSCHAFT,
GESCHICHTSWISSENSCHAFTEN,
HISTORISCH,
HISTORISCHE,
HISTORISCHEN,
HISTORISCHER,
HISTORISCHERN,
GESCHICHTLICH,
GESCHICHTLICHE,
GESCHICHTLICHEN,
GESCHICHTLICHER,
GESCHICHTLICHERN,
LEGENDE,
LEGENDEN,
MYTHUS,
MYTHOS,
MYTHE,
MYTHEN,
MYTHERN,
SAGE,
SAGEN,
MUSEUM,
MUSEEN,
MYTHOLOGIE,
MYTHOLOGIEN,
KODEX,
KODEXEN,
MONOLITHE,
MONOLITHEN,
TEMPEL,
TEMPLEN,
#

<Aztec_culture.syn.Ger> 6

AZTEKISCHE ERBSCHAFT,
AZTEKISCHE KULTUR,
AZTEKISCH EINFLUB,
AZTEKISCHER EINFLUSSE,
AZTEKISCHE GESCHICHTE,
AZTEKISCHE GESCHICHTSWISSENSHAFT,
AZTEKISCHER TEMPEL,
AZTEKISCH TEMPEL,
AZTEKISCH MYTHOS,
AZTEKISCH MYTHUS,
AZTEKISCHE MYTHE,
AZTEKISCHER MYTHEN,
AZTEKISCHE LEGENDE,
AZTEKISCHEN LEGENDEN,
AZTEKISCHE MYTHOLOGIE,
AZTEKISCHEN MYTHOLOGIEN,
AZTEKISCH KODEX,
AZTEKISCHER KODEXE,
AZTEKISCH GOTT,
AZTEKISCHER GOTTER,
AZTEKISCHERN GOTTERN,
AZTEKISCHE GOTTIN,
AZTEKISCHEN GOTTINEN,
AZTEKISCHE SAGE,
AZTEKISCHEN SAGEN,
#