The Effects of Language Relatedness on Multilingual Information Retrieval: A Case Study With Indo-European and Semitic Languages

Peter A. Chew
Sandia National Laboratories
P. O. Box 5800, MS 1012
Albuquerque, NM 87185-1012, USA
pchew@sandia.gov

Ahmed Abdelali
New Mexico State University
P.O. Box 30002, Mail Stop 3CRL
Las Cruces, NM 88003-8001, USA
ahmed@crl.nmsu.edu

Abstract

We explore the effects of language relatedness within a multilingual information retrieval (IR) framework which can be deployed to virtually any language, focusing specifically on Indo-European versus Semitic languages. The Semitic languages present unique challenges to IR for a number of reasons, so we set out to answer the question of whether cross-language IR for Semitic languages can be boosted by manipulation of the training data (which, in our framework, includes multilingual parallel text, some of which is morphologically analyzed). We attempted three measures to achieve this: first, the inclusion of genetically related (i.e., other Semitic) languages in the training data; second, the inclusion of non-related languages sharing the same script, and third, the inclusion of morphological analysis for Semitic languages. We find that language relatedness is a definite factor in boosting IR precision; script similarity can probably be ruled out as a factor; and morphological analysis can be helpful, but – perhaps paradoxically – not necessarily to the languages which are subjected to morphological analysis.

1 Introduction

In this paper, we consider how related languages fit into a general framework developed for multilingual cross-language information retrieval (CLIR). Although this framework can deal with virtually any language, there are some special considerations which make related languages more interesting for exploration. Taking one example, Semitic languages are distinguished by their complex morphology, a characteristic which presents challenges to an information retrieval model in which terms (usually, separated by white space or punctuation) are implicitly treated as individual units of meaning. We consider three possible methods for investigating the phenomena. In all cases, we keep the overall framework the same but simply make changes to the training data.

One method we consider is to augment the training data with text from related languages; we compare results obtained from using Semitic languages with those obtained when non-Semitic languages are used. The other two relate to morphological analysis: the second is to replace inflected forms (in just one language, Arabic) with just the root in the training data; and the third is to remove vowels (again in just one language, Hebrew).

The paper is organized as follows. Section 2 describes our general framework, which is a standard one used for CLIR. At a high level, section 3 outlines some of the challenges Semitic languages present within the context of our approach. In section 4, we compare results from using a number of different combinations of training data with the same test data. Finally, we conclude on our findings in section 5.

2 The Framework

2.1 General description

The framework that we use for IR is multilingual Latent Semantic Analysis (LSA) as described by Berry et al. (1994:21, and used by Landauer and Littman (1990) and Young (1994). A number of different approaches to CLIR have been proposed; generally, they rely either on the use of a parallel

corpus for training, or translation of the IR query. Either or both of these methods can be based on the use of dictionaries, although that is not the approach that we use.

In the standard multilingual LSA framework, a term-by-document matrix is formed from a parallel aligned corpus. Each 'document' consists of the concatenation of all the languages, so terms from all languages will appear in any given document. Thus, if there are K languages, N documents (each of which is translated into each of the K languages), and T distinct linguistic terms across all languages, then the term-by-document matrix is of dimensions T by N. Each cell in the matrix represents a weighted frequency of a particular term *t* (in any language) in a particular document *n*. The weighting scheme we use is a standard log-entropy scheme in which the weighted frequency $x_{t,n}$ of a particular term *t* in a particular document *n* is given by:

$$W = \log_2 (F + 1) \times (1 + H_t / \log_2 (N))$$

where F is the raw frequency of *t* in *n*, and H_t is a measure of the entropy of the term across all documents. The last term in the expression above, $\log_2 (N)$, is the maximum entropy that any term can have in the corpus, and therefore $(1 + H_t / \log_2 (N))$ is 1 for the most distinctive terms in the corpus, 0 for those which are least distinctive. The log-entropy weighting scheme has been shown to outperform other schemes such as tf-idf in LSA-based retrieval (see for example Dumais 1991).

The sparse term-by-document matrix is subjected to singular value decomposition (SVD), and a reduced non-sparse matrix is output. Generally, we used the output corresponding to the top 300 singular values in our experiments.

To evaluate the similarity of unseen queries or documents (those not in the training set) to one another, these documents are tokenized, the weighted frequencies are calculated in the same way as they were for the training set, and the results are multiplied by the matrices output by the SVD to project the unseen queries/documents into a 'semantic space', assigning (in our case) 300-dimensional vectors to each document. Again, our approach to measuring the similarity of one document to another is a standard one: we calculate the cosine between the respective vectors.

For CLIR, the main advantages of an approach like LSA are that it is by now quite well-understood; the underlying algorithms remain constant regardless of which languages are being compared; and there is wide scope to use different sets of training data, providing they exist in parallel corpora. LSA is thus a highly generic approach to CLIR: since it relies only on the ability to tokenize text at the boundaries between words, or more generally semantic units, it can be generalized to virtually all languages.

2.2 Training and test data

For our experiments, the training and test data were taken from the Bible and Quran respectively. As training data, the Bible lends itself extremely well to multilingual LSA. It is highly available in multiple languages[1] (over 80 parallel translations in 50 languages, mostly public-domain, are available from a single website, www.unboundbible.org); and a very fine-grained alignment is possible (by verse) (Resnik et al 1999, Chew and Abdelali 2007). Many purpose-built parallel corpora are biased towards particular language groups (for example, the European Union funds work in CLIR, but it tends to be biased towards European languages – for example, see Peters 2001). This is not as true of the Bible, and the fact that it covers a wider range of languages is a reflection of the reasons it was translated in the first place.

The question which is most commonly raised about use of the Bible in this way is whether its coverage of vocabulary from other domains is sufficient to allow it to be used as training data for most applications. Based on a variety of experiments we have carried out (see for example Chew et al. forthcoming), we believe this need not always be a drawback – it depends largely on the intended application. However, it is beyond our scope to address this in detail here; it is sufficient to note that for the experiments we describe in this paper, we were able to achieve perfectly respectable CLIR results using the Bible as the training data.

[1] It has proved hard to come by reliable statistics to allow direct comparison, but the Bible is generally believed to be the world's most widely translated book. At the end of 2006, it is estimated that there were full translations into 429 languages and partial translations into 2,426 languages (Bible Society 2007).

As test data, we used the 114 suras (chapters) of the Quran, which has also been translated into a wide variety of languages. Clearly, both training and test data have to be available in multiple languages to allow the effectiveness of CLIR to be measured in a meaningful way. For the experiments reported in this paper, we limited the testing languages to Arabic, English, French, Russian and Spanish (the respective abbreviations AR, EN, FR, RU and ES are used hereafter). The test data thus amounted to 570 (114 × 5) documents: a relatively small set, but large enough to achieve statistically significant results for our purposes, as will be shown. In all tests described in this paper, we use the same test set: thus, although the test documents all come from a single domain, it is reasonable to suppose that the comparative results can be generalized to other domains.

The complete list of languages used for both testing and training is given in Table 1.

Language	Bible -training-	Quran -test-	Language Family	Sub-Family
Afrikaans	Yes	No	Indo-European	Germanic-West
Amharic	Yes	No	Afro-Asiatic	Semitic-South
Arabic	Yes	Yes	Afro-Asiatic	Semitic-Central
Aramaic	Yes	No	Afro-Asiatic	Semitic-North
Czech	Yes	No	Indo-European	Slavic-West
Danish	Yes	No	Indo-European	Germanic-North
Dutch	Yes	No	Indo-European	Germanic-West
English	Yes	Yes	Indo-European	Germanic-West
French	Yes	Yes	Indo-European	Italic
Hebrew	Yes	No	Afro-Asiatic	Semitic-Central
Hungarian	Yes	No	Uralic	Finno-Ugric
Japanese	Yes	No	Altaic	
Latin	Yes	No	Indo-European	Italic
Persian	Yes	No	Indo-European	Indo-Iranian
Russian	Yes	Yes	Indo-European	Slavic-East
Spanish	Yes	Yes	Indo-European	Italic

Table 1. Languages used for training and testing

2.3 Test method

We tokenized each of the 570 test documents, applying the weighting scheme described above to obtain a vector of weighted frequencies of each term in the document, then multiplying that vector by $U \times S^{-1}$, also as described above. The result was a set of projected document vectors in the 300-dimensional LSA space.

For some of our experiments, we used a light stemmer for Arabic (Darwish 2002) to replace inflected forms in the training data with citation forms. It is commonly accepted that morphology improves IR (Abdou et al. 2005, Lavie et al. 2004, Larkey et al. 2002, Oard and Gey 2002), and it will be seen that our results generally confirm this.

For Hebrew, we used the Westminster Leningrad Codex in the training data. Since this is available for download either with vowels or without vowels, no morphological pre-processing was required in this case; we simply substituted one version for the other in the training data when necessary.

Various measurements are used for evaluating IR systems performance (Van Rijsbergen 1979). However, since the aim of our experiments is to assess whether we could identify the correct translation for a given document among a set of possibilities in another language (i.e., given the language of the query and the language of the results), we selected 'precision at 1 document' as our preferred metric. This metric represents the proportion of cases, on average, where the translation was retrieved first.

3 Challenges of Semitic languages

The features which make Semitic languages challenging for information retrieval are generally fairly well understood: it is probably fair to say that chief among them is their complex morphology (for example, ambiguity resulting from diacritization, root-and-pattern alternations, and the use of infix morphemes as described in Habash 2004).

These challenges can be illustrated by means of a statistical comparison of a portion of our training data (the Gospel of Matthew) as shown in Table 2.

	Types	Tokens
Afrikaans	2,112	24,729
French	2,840	24,438
English	2,074	23,503
Dutch	2,613	23,099
Danish	2,649	21,816
Spanish	3,075	21,279
Persian	3,587	21,190
Hungarian	4,730	18,787
Czech	4,236	18,000
Russian	4,196	16,826
Latin	3,936	16,543
Hebrew (Modern)	4,337	14,153
Arabic	4,607	13,930
Japanese	5,741	13,130
Amharic	5,161	12,940
TOTAL	55,894	284,363

Table 2. Statistics of parallel texts by language

From Table 2, it should be clear that there is generally an inverse relationship between the number of types and tokens. Modern Indo-European (IE) (and particularly Germanic or Italic languages) are at one end of the spectrum, while the Semitic languages (along with Japanese) are at the other. The statistics separate 'analytic' languages from 'synthetic' ones, and essentially illustrate the fact that, thanks to the richness of their morphology, the Semitic languages pack more information (in the information-theoretic sense) into each term than the other languages. Because this results in higher average entropy per word (in the information theoretic sense), a challenge is presented to information retrieval techniques such as LSA which rely on tokenization at word boundaries: it is harder to isolate each 'unit' of meaning in a synthetic language. The actual effect this has on information retrieval precision will be shown in the next section.

4 Results with LSA

The series of experiments described in this section have the aims of:

- clarifying what effect morphological analysis of the training data has on CLIR precision;
- highlighting the effect on CLIR precision of adding more languages in training;
- illustrating what the impact is of adding a partial translation (text in one language which is only partially parallel with the texts in the other languages)

We choose Arabic as the language of focus in our experiment; specifically for these experiments, we intended to reveal the effect of adding languages from the same group (Semitic) compared with that of adding languages of different groups.

First, we present results in Table 3 which confirm that morphological analysis of the training data improves CLIR performance.

	ES	RU	FR	EN	AR
without morphological analysis of Arabic					
ES	1.0000	0.5614	0.8333	0.7368	0.2895
RU	0.4211	1.0000	0.5263	0.7632	0.2632
FR	0.7807	0.7018	1.0000	0.8158	0.4035
EN	0.7193	0.8158	0.8596	1.0000	0.4825
AR	0.5000	0.2807	0.6228	0.5526	1.0000
Average precision: Overall 0.677, within IE 0.783, IE-Semitic 0.488					
with morphological analysis of Arabic					
ES	1.0000	0.6579	0.8772	0.7807	0.4123
RU	0.4912	1.0000	0.7193	0.8158	0.3947
FR	0.8421	0.7719	1.0000	0.8421	0.3772
EN	0.8070	0.8684	0.8947	1.0000	0.3684
AR	0.3947	0.3509	0.5614	0.4561	1.0000
Average precision: Overall 0.707, within IE 0.836, IE-Semitic 0.480					

Table 3. Effect of morphological analysis[2]

An important point to note first is that CLIR precision is generally much lower for pairs including Arabic than it is elsewhere, lending support to our assertion above that Arabic and other Semitic languages present special challenges in information retrieval.

It also emerges from Table 3 that when morphological analysis of Arabic was added, the overall average precisions increased from 0.677 to 0.707, a highly significant increase ($p \approx 6.7 \times 10^{-8}$). (Here and below, a chi-squared test is used to measure statistical significance.)

Given that the ability of morphological analysis to improve IR precision has been documented, this result in itself is not surprising. However, it is interesting that the net benefit of adding morphological analysis – and just to Arabic within the training data – was more or less confined to pairs of non-Semitic languages. We believe that the explanation is that by adding morphology more relations (liai-

[2] In this and the following tables, the metric used is precision at 1 document (discussed in section 2.3).

sons) are defined in LSA between the words from different languages. For language pairs including Arabic, the average precision actually decreased from 0.488 to 0.480 when morphology was added (although this decrease is insignificant).

With the same five training languages as used in Table 3, we added Persian. The results are shown in Table 4.

	ES	RU	FR	EN	AR
ES	1.0000	0.6140	0.8246	0.7632	0.3246
RU	0.5088	1.0000	0.6667	0.7982	0.2281
FR	0.8772	0.7368	1.0000	0.8158	0.3947
EN	0.8246	0.8333	0.8947	1.0000	0.4035
AR	0.4474	0.4386	0.6140	0.5526	1.0000
Average precision: Overall 0.702, within IE 0.822, IE-Semitic 0.489					

Table 4. Effect on CLIR of adding Persian

First to note is that the addition of Persian (an IE language) led to a general increase in precision for pairs of IE languages (Spanish, Russian, French and English) from 0.783 to 0.822 but no significant change for pairs including Arabic (0.488 to 0.489). Although Persian and Arabic share the same script, these results confirm that genetic relatedness is a much more important factor in affecting precision.

Chew and Abdelali (2007) show that the results of multilingual LSA generally improve as the number of parallel translations used in training increases. Our next step here, therefore, is to analyze whether it makes any difference whether the additional languages are from the same or different language groups. In Table 5 we compare the results of adding an IE language (Latin), an Altaic language (Japanese), and another Semitic language (Hebrew) to the training data. In all three cases, no morphological analysis of the training data was performed.

Based on these results, cross-language precision yielded only very slightly improved results overall by adding Latin or Japanese. With Japanese, the net improvement (0.677 to 0.680) was not statistically significant overall, neither was the change significant for pairs either including or excluding Arabic (0.488 to 0.485 and 0.783 to 0.789 respectively). Note that this is even though Japanese shares some statistical (although of course not linguistic) properties with the Semitic languages, as shown in Table 2. With Latin, the net overall improvement (0.677 to 0.699) was barely significant ($p \approx 0.01$) and was insignificant for pairs including Arabic (0.488 to 0.496). With Hebrew, however, the net improvement was highly significant in all cases (0.677 to 0.718, $p \approx 3.36 \times 10^{-6}$ overall, 0.783 to 0.819, $p \approx 2.20 \times 10^{-4}$ for non-Semitic pairs, and 0.488 to 0.538, $p \approx 1.45 \times 10^{-3}$ for pairs including Arabic). We believe that these results indicate that there is more value overall in ensuring that languages are paired with at least one other related language in the training data; our least impressive results (with Japanese) were when two languages in training (one Semitic and one Altaic language) were 'isolated'.

	ES	RU	FR	EN	AR
Latin included in training data					
ES	1.0000	0.6140	0.8333	0.7456	0.2544
RU	0.4737	1.0000	0.6316	0.8246	0.3333
FR	0.8596	0.7368	1.0000	0.8333	0.4474
EN	0.7719	0.7982	0.8860	1.0000	0.4474
AR	0.5088	0.3509	0.6140	0.5088	1.0000
Average precision: Overall 0.699, within IE 0.813, IE-Semitic 0.496					
Japanese included in training data					
ES	1.0000	0.5789	0.8333	0.7456	0.2895
RU	0.4298	1.0000	0.5526	0.7807	0.2719
FR	0.7719	0.7368	1.0000	0.8070	0.4035
EN	0.7193	0.807	0.8596	1.0000	0.4123
AR	0.5088	0.2982	0.614	0.5702	1.0000
Average precision: Overall 0.680, within IE 0.789, IE-Semitic 0.485					
Modern Hebrew (no vowels) in training data					
ES	1.0000	0.6140	0.8596	0.7807	0.3509
RU	0.4561	1.0000	0.6667	0.7719	0.3684
FR	0.8509	0.7193	1.0000	0.8684	0.4298
EN	0.7632	0.8509	0.9035	1.0000	0.4298
AR	0.5263	0.4474	0.6491	0.6404	1.0000
Average precision: Overall 0.718, within IE 0.819, IE-Semitic 0.538					

Table 5. Effect of language relatedness on CLIR

The next set of results are for a repetition of the previous three experiments, but this time with morphological analysis of the Arabic data. These results are shown in Table 6.

As was the case without the additional languages, the overall effect of adding morphological analysis of Arabic is still to increase precision. In all three cases, the net improvement for pairs excluding Arabic is highly significant (0.813 to 0.844 with Latin, 0.789 to 0.852 with Japanese, and 0.819 to 0.850 with Hebrew). For pairs including Arabic, however, the change is again insignificant. This was a consistent but surprising feature of our results, that morphological analysis of Arabic in fact appears to benefit non-Semitic languages more

than it benefits Arabic itself, at least with this dataset. The results might possibly have been different if we had included other Semitic languages in the test data, although this appears unlikely as we found the same phenomenon consistently occurring across a wide variety of tests, and regardless of which languages we used in training.

	ES	RU	FR	EN	AR
Latin included in training data					
ES	1.0000	0.6579	0.8684	0.7456	0.4211
RU	0.5614	1.0000	0.7456	0.8509	0.4386
FR	0.8421	0.8158	1.0000	0.8509	0.4211
EN	0.8421	0.8333	0.8947	1.0000	0.4123
AR	0.4123	0.3947	0.5351	0.4825	1.0000
Average precision: Overall 0.721, within IE 0.844, IE-Semitic 0.502					
Japanese included in training data					
ES	1.0000	0.7544	0.8684	0.8070	0.4211
RU	0.4737	1.0000	0.7193	0.8509	0.4123
FR	0.8246	0.8596	1.0000	0.8772	0.4211
EN	0.8421	0.8596	0.8947	1.0000	0.4035
AR	0.3333	0.3509	0.5614	0.4649	1.0000
Average precision: Overall 0.720, within IE 0.852, IE-Semitic 0.485					
Modern Hebrew (no vowels) in training data					
ES	1.0000	0.7018	0.9035	0.7982	0.4561
RU	0.5614	1.0000	0.7105	0.8070	0.4035
FR	0.8421	0.8246	1.0000	0.8596	0.4825
EN	0.8509	0.8509	0.8947	1.0000	0.4123
AR	0.3947	0.4298	0.5351	0.5175	1.0000
Average precision: Overall 0.729, within IE 0.850, IE-Semitic 0.514					

Table 6. Effect of language relatedness and morphology on CLIR

For further verification, we explored what would happen if only the Arabic root were included in morphological analysis. As already mentioned, for languages that combine affixes with the stem, there is a higher token-to-type ratio. Omitting the affix from the morphological analysis of these languages reveals the importance of considering the affixes and their contribution to the semantics of a given sentence. Although LSA is not sentence-structure-aware (as it uses a bag-of-words approach), the importance of considering the affixes as part of the sentence is very crucial. The results in Table 7 demonstrate clearly that ignoring or over-looking the word affixes has a negative effect on the overall performance of the CLIR system. When including only the Arabic stem, a performance degradation is noticeable across all languages, with a larger impact on IE languages. The results which illustrate can be seen by comparing Table 7 with Table 3.

	ES	RU	FR	EN	AR
morphological analysis of Arabic –Stem only-					
ES	1.0000	0.5789	0.8070	0.7807	0.3421
RU	0.4912	1.0000	0.6842	0.8246	0.1842
FR	0.8421	0.7018	1.0000	0.8333	0.4211
EN	0.8333	0.8333	0.9211	1.0000	0.4211
AR	0.4561	0.4386	0.5702	0.4912	1.0000
Average precision: Overall 0.698, within IE 0.821, IE-Semitic 0.481					

Table 7. Effect of Using Stem only

Next, we turn specifically to a comparison of the effect that different Semitic languages have on CLIR precision. Here, we compare the results when the sixth language used in training is Hebrew, Amharic, or Aramaic. However, since our Amharic and Aramaic training data were only partially parallel (we have only the New Testament in Amharic, and only portions of the New Testament in Aramaic), we first considered the effect that partial translations have on precision. Table 8 shows the results we obtained when only the Hebrew Old Testament (with vowels) was used as the sixth parallel version. No morphological analysis was performed.

	ES	RU	FR	EN	AR
without morphological analysis of Arabic					
ES	1.0000	0.6842	0.8421	0.8158	0.3947
RU	0.4211	1.0000	0.6228	0.7982	0.4737
FR	0.8509	0.7719	1.0000	0.8509	0.4737
EN	0.7895	0.8333	0.8684	1.0000	0.4649
AR	0.4561	0.3333	0.6404	0.4561	1.0000
Average precision: Overall 0.714, within IE 0.822, IE-Semitic 0.521					
with morphological analysis of Arabic					
ES	1.0000	0.7105	0.9035	0.8333	0.4737
RU	0.4649	1.0000	0.7456	0.8333	0.4912
FR	0.8421	0.8070	1.0000	0.8860	0.4474
EN	0.8772	0.8421	0.9298	1.0000	0.4298
AR	0.2719	0.3684	0.5088	0.5000	1.0000
Average precision: Overall 0.727, within IE 0.855, IE-Semitic 0.499					

Table 8. Effect of partial translation on CLIR

Although two or more parameters differ from those used for Hebrew in Table 5 (a fully-parallel text in modern Hebrew without vowels, versus a partial text in Ancient Hebrew with vowels), it is worth comparing the two sets of results. In particular, the reductions in average precision from 0.718 to 0.714 and from 0.729 to 0.727 respectively are

insignificant. Likewise, the changes for pairs with and without Arabic were insignificant. This appears to show that, at least up to a certain point, even only partially parallel corpora can successfully be used under our LSA-based approach. We now turn to the results we obtained using Aramaic, with the intention of comparing these to our previous results with Hebrew.

	ES	RU	FR	EN	AR
no morphological analysis of Arabic					
ES	1.0000	0.4035	0.8070	0.7368	0.2632
RU	0.3509	1.0000	0.5965	0.6579	0.2281
FR	0.8421	0.6754	1.0000	0.8246	0.2719
EN	0.7018	0.6754	0.8947	1.0000	0.2719
AR	0.4825	0.2807	0.4649	0.3947	1.0000
Average precision: Overall 0.633, within IE 0.760, IE-Semitic 0.406					
morphological analysis of Arabic					
ES	1.0000	0.5351	0.8684	0.7719	0.2895
RU	0.5175	1.0000	0.6930	0.7807	0.3421
FR	0.8947	0.7807	1.0000	0.8684	0.2807
EN	0.8070	0.8158	0.9035	1.0000	0.2982
AR	0.3509	0.2193	0.3772	0.2895	1.0000
Average precision: Overall 0.667, within IE 0.827, IE-Semitic 0.383					

Table 9. Effect of Aramaic on CLIR

Here, there is a noticeable across-the-board decrease in precision from the previous results. We believe that this may have more to do with the fact that the Aramaic training data we have is fairly sparse (2,957 verses of the Bible out of a total of 31,226, compared with 23,269 out of 31,226 for Ancient Hebrew). It is likely that at some point as the parallel translation's coverage drops (somewhere between the coverage of the Hebrew and the Aramaic), there is a severe hit to the performance of CLIR. Accordingly, we discarded Aramaic for further tests.

Next, we considered the addition of two Semitic languages other than Arabic, Modern Hebrew and Amharic, to the training data. In this case, we performed morphological analysis of Arabic.

The results appear to show a significant increase in precision for pairs of IE languages and a significant *decrease* for cross-language-group cases (those where an IE language is paired with Arabic), compared to when just Modern Hebrew was used in the training data (see the relevant part of Table 6). It is not clear why this is the case, but in this case we believe that it is quite possible that the results would have been different if more than one Semitic language had been included in the test data.

	ES	RU	FR	EN	AR
ES	1.0000	0.6930	0.8860	0.7719	0.4649
RU	0.5000	1.0000	0.7456	0.8684	0.5175
FR	0.8772	0.7982	1.0000	0.8772	0.4649
EN	0.8684	0.8596	0.9298	1.0000	0.4386
AR	0.2632	0.2982	0.4386	0.3947	1.0000
Average precision: Overall 0.718, within IE 0.855, IE-Semitic 0.476					

Table 10. CLIR with 7 languages (including Modern Hebrew and Amharic)

We now come to a rare example where we achieved a boost in precision specifically for Arabic. In this case, we repeated the last experiment but removed the vowels from the Hebrew text. The results are shown in Table 11.

	ES	RU	FR	EN	AR
ES	1.0000	0.7018	0.8772	0.8158	0.5088
RU	0.5175	1.0000	0.7632	0.8421	0.4825
FR	0.8596	0.8246	1.0000	0.8860	0.5351
EN	0.8947	0.8158	0.9298	1.0000	0.5088
AR	0.2895	0.3772	0.5526	0.5000	1.0000
Average precision: Overall 0.739, within IE 0.858, IE-Semitic 0.528					

Table 11. Effect of removing Hebrew vowels

Average precision for pairs including Arabic increased from 0.476 to 0.528, an increase which was significant ($p \approx 7.33 \times 10^{-4}$), but for other pairs the change was insignificant. Since the Arabic text in training did not include vowels, we believe that the exclusion of vowels from Hebrew placed the two languages on a more common footing, allowing LSA, for example, to make associations between Hebrew and Arabic roots which otherwise might not have been made. Although Hebrew and Arabic do not always share common stems, it can be seen from Table 2 that the type/token statistics of Hebrew (without vowels) and Arabic are very similar. The inclusion of Hebrew vowels would change the statistics for Hebrew considerably, increasing the number of types (since previously indistinguishable wordforms would now be listed separately). Thus, with the *exclusion* of Hebrew vowels, there should be more instances where Arabic tokens can be paired one-to-one with Hebrew tokens.

Finally, in order to confirm our conclusions and to eliminate any doubts about the results obtained so far, we experimented with more languages. We added Japanese, Afrikaans, Czech, Danish, Dutch,

Hungarian and Hebrew in addition to our 5 original languages. Morphological analysis of the Arabic text in training was performed, as in some of the previous experiments. The results of these tests are shown in Table 12.

	ES	RU	FR	EN	AR
11 languages (original 5 + Japanese, Afrikaans, Czech, Danish, Dutch, and Hungarian)					
ES	1.0000	0.6754	0.9035	0.7719	0.5526
RU	0.4737	1.0000	0.7632	0.8772	0.5175
FR	0.8596	0.8070	1.0000	0.8947	0.5088
EN	0.8421	0.8684	0.9035	1.0000	0.4912
AR	0.3772	0.2632	0.6316	0.4912	1.0000
Average precision: Overall 0.739, within IE 0.853, IE-Semitic 0.537					
12 languages (as above plus Hebrew)					
ES	1.0000	0.7018	0.8947	0.7719	0.6404
RU	0.6667	1.0000	0.7105	0.9123	0.6228
FR	0.8772	0.8333	1.0000	0.8421	0.6404
EN	0.6667	0.8684	0.9035	1.0000	0.6316
AR	0.5877	0.4386	0.5965	0.6491	1.0000
Average precision: Overall 0.778, within IE 0.853, IE-Semitic 0.645					

Table 12. Effect of further languages on CLIR

Generally, these results confirm the finding of Chew and Abdelali (2007) about adding more languages; doing so enhances the ability to identify translations across language boundaries. Across the board (for Arabic and other languages), the increase in precision gained by adding Afrikaans, Czech, Danish, Dutch and Hungarian is highly significant (compared to the part of Table 5 which deals with Japanese, overall average precision increased from 0.680 to 0.739, with $p \approx 1.17 \times 10^{-11}$; for cross-language-group retrieval, from 0.485 to 0.537, with $p \approx 9.31 \times 10^{-4}$; for pairs within IE, from 0.789 to 0.853 with $p \approx 2.81 \times 10^{-11}$). In contrast with most previous results, however, with the further addition of Hebrew, precision was boosted primarily for Arabic (0.537 to 0.645 with $p \approx 4.39 \times 10^{-13}$). From this and previous results, it appears that there is no clear pattern to when the addition of a Semitic language in training was beneficial to the Semitic language in testing.

5 Conclusion and future work

Based on our results, it appears that although clear genetic relationships exist between certain languages in our training data, it was less possible than we had anticipated to leverage this to our advantage. We had expected, for example, that by including multiple Semitic languages in the training data within an LSA framework, we would have been able to improve cross-language information retrieval results specifically for Arabic. Perhaps surprisingly, the greatest benefit of including additional Semitic languages in the training data is most consistently to non-Semitic languages. A clear observation is that *any* additional languages in training are generally beneficial, and the benefit of additional languages can be considerably greater than the benefits of linguistic pre-processing (such as morphological analysis). Secondly, it is not necessarily the case that cross-language retrieval with Arabic is helped most by including other Semitic languages, despite the genetic relationship. Finally, as we expected, we were able to rule out script similarity (e.g. between Persian and Arabic) as a factor which might improve precision. Our results appear to demonstrate clearly that language relatedness is much more important in the training data than use of the same script.

Finally, to improve cross-language retrieval with Arabic – the most difficult case in the languages we tested – we attempted to 'prime' the training data by including Arabic morphological analysis. This did lead to a statistically significant improvement overall in CLIR, but – perhaps paradoxically – the improvement specifically for cross-language retrieval with Arabic was negligible in most cases. The only two measures which were successful in boosting precision for Arabic significantly were (1) the inclusion of Modern Hebrew in the training data; and (2) the elimination of vowels in the Ancient Hebrew training data – both measures which would have placed the training data for the two Semitic languages (Arabic and Hebrew) on a more common statistical footing. These results appear to confirm our hypothesis that there is value, within the current framework, of 'pairing' genetically related languages in the training data. In short, language relatedness does matter in cross-language information retrieval.

6 Acknowledgement

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

7 References

Abdou, S., Ruck, P., and Savoy, J. 2005. Evaluation of Stemming, Query Expansion and Manual Indexing Approaches for the Genomic Task. In *Proceedings of TREC 2005*.

Berry, M. W., Dumais, S. T., and O'Brien, G. W. 1994. Using Linear Algebra for Intelligent Information Retrieval. *SIAM: Review*, 37, 573-595.

Biola University. 2005-2006. *The Unbound Bible*. Accessed at http://www.unboundbible.com/ on February 27, 2007.

Chew, P. A., and Abdelali, A. 2007. *Benefits of the 'Massively Parallel Rosetta Stone': Cross-Language Information Retrieval with over 30 Languages*, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, ACL 2007. Prague, Czech Republic, June 23–30, 2007. pp. 872-879.

Chew, P. A., Kegelmeyer, W. P., Bader, B. W. and Abdelali, A. Forthcoming. *The Knowledge of Good and Evil: Multilingual Ideology Classification with PARAFAC2 and Maching Learning*.

Chew, P. A., Verzi, S. J., Bauer, T. L., and McClain, J. T. 2006. Evaluation of the Bible as a Resource for Cross-Language Information Retrieval. *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*, 68–74.

Darwish, K. 2002. *Building a shallow Arabic morphological analyzer in one day*. In Proceedings of the Association for Computational Linguistics (ACL-02), 40th Anniversary Meeting. pp. 47-54.

Dumais, S. T. 1991. Improving the Retrieval of Information from External Sources. *Behavior Research Methods, Instruments, and Computers* 23 (2), 229-236.

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S. and Harshman, R. 1998. Using Latent Semantic Analysis to Improve Access to Textual Information. In *CHI'88: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 281-285. ACM Press.

Frakes, W. B. and Baeza-Yates, R. 1992. *Information Retrieval: Data Structures and Algorithms*. Prentice-Hall: New Jersey.

Habash, N. 2004. Large Scale Lexeme Based Arabic Morphological Generation. In *Proc. of Traitement Automatique du Langage Naturel.*

Larkey, L., Ballesteros, L. and Connell, M. 2002. Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-Occurrence Analysis. *SIGIR 2002*, Finland, pp. 275-282.

Larkey, L. and Connell, M. 2002. Arabic Information Retrieval at Umass in TREC-10. In Voorhees, E.M. and Harman, D.K. (eds.): *The Tenth Text Retrieval Conference, TREC 2001 NIST Special Publication 500-250*, pp. 562-570.

Lavie, A., Peterson, E., Probst, K., Wintner, S., and Eytani, Y. 2004. Rapid Prototyping of a Transfer-Based Hebrew-to-English Machine Translation System. In *Proceedings of the TMI-04*.

Mathieu, B., Besançon, R. and Fluhr, C. 2004. Multilingual Document Clusters Discovery. *Recherche d'Information Assistée par Ordinateur (RIAO) Proceedings*, 1-10.

Oard, D. and Gey, F. 2002. *The TREC 2002 Arabic/English CLIR Track, NIST TREC 2002 Proceedings*, pp. 16-26.

Peters, C. (ed.). 2001. *Cross-Language Information Retrieval and Evaluation: Workshop of the Cross-Language Evaluation Forum, CLEF 2000*. Berlin: Springer-Verlag.

Resnik, P., Olsen, M. B., and Diab, M. 1999. The Bible as a Parallel Corpus: Annotating the "Book of 2000 Tongues". *Computers and the Humanities*, 33, 129-153.

Van Rijsbergen, C. 1979. *Information Retrieval (2nd edition)*. Butterworth: London.