# Benefits of the ‘Massively Parallel Rosetta Stone’: Cross-Language Information Retrieval with over 30 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

In this paper, we describe our experiences in extending a standard cross-language information retrieval (CLIR) approach which uses parallel aligned corpora and Latent Semantic Indexing. Most, if not all, previous work which follows this approach has focused on bilingual retrieval; two examples involve the use of French-English or English-Greek parallel corpora. Our extension to the approach is ‘massively parallel’ in two senses, one linguistic and the other computational. First, we make use of a parallel aligned corpus consisting of almost 50 parallel translations in over 30 distinct languages, each in over 30,000 documents. Given the size of this dataset, a ‘massively parallel’ approach was also necessitated in the more usual computational sense. Our results indicate that, far from adding more noise, more linguistic parallelism is better when it comes to cross-language retrieval precision, in addition to the self-evident benefit that CLIR can be performed on more languages.

## 1 Introduction

Approaches to cross-language information retrieval (CLIR) fall generally into one of two types, or some combination thereof: the ‘query translation’ approach or the ‘parallel corpus’ approach. The first of these, which is perhaps more common, involves translation of the query into the target language, for example using machine translation or on-line dictionaries. The second makes use of parallel aligned corpora as training sets. One approach which uses parallel corpora does this in conjunction with Latent Semantic Indexing (LSI) (Landauer and Littman 1990, Young 1994). According to Berry et al. (1994:21), the use of LSI with parallel corpora can be just as effective as the query translation approach, and avoids some of the drawbacks of the latter, discussed in Nie et al. (1999).

Generally, research in CLIR has not attempted to use very many languages at a time (see for example Nie and Jin 2002). With query translation (although that is not the approach that Nie and Jin take), this is perhaps understandable, as for each new language, a new translation algorithm must be included. The effort involved in extending query translation to multiple languages, therefore, is likely to be in proportion to the number of languages.

With parallel corpora, the reason that research has been limited to only a few languages at a time – and usually just two at a time, as in the LSI work cited above – is more likely to be rooted in the widespread perception that good parallel corpora are difficult to obtain (see for example Asker 2004). However, recent work (Resnik et al. 1999, Chew et al. 2006) has challenged this idea.

One advantage of a ‘massively parallel’ multilingual corpus is perhaps self-evident: within the LSI framework, the more languages are mapped into the single conceptual space, the fewer restrictions there are on which languages documents can be selected from for cross-language retrieval. However, several questions were raised for us as



*Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 872–879, Prague, Czech Republic, June 2007. 

we contemplated the use of a massively parallel corpus. Would the addition of languages not used in testing create 'noise' for a given language pair, reducing the precision of CLIR? Could *partially* parallel corpora be used? Our work appears to show both that more languages are generally beneficial, and even incomplete parallel corpora can be used. In the remainder of this paper, we provide evidence for this claim. The paper is organized as follows: section 2 describes the work we undertook to build the parallel corpus and its characteristics. In section 3, we outline the mechanics behind the 'Rosetta-Stone' type method we use for cross-language comparison. In section 4, we present and discuss the results of the various tests we performed. Finally, we conclude on our findings in section 5.

## 2 The massively parallel corpus

Following Chew et al. (2006), our parallel corpus was built up from translations of the Bible which are freely available on the World Wide Web. Although reliable comparable statistics are hard to find, it appears to be generally agreed that the Bible is the world's most widely translated book, with complete translations in 426 languages and partial translations in 2,403 as of December 31, 2005 (Bible Society, 2006). Great care is taken over the translations, and they are alignable by chapter and verse. According to Resnik et al. (1999), the Bible's coverage of modern vocabulary may be as high as 85%. The vast majority of the translations we used came from the 'Unbound Bible' website (Biola University, 2005-2006); from this website, the text of a large number of different translations of the Bible can – most importantly for our purposes – be downloaded in a tab-delimited format convenient for loading into a database and then indexing by chapter and verse in order to ensure 'parallelism' in the corpus. The number of translations available at the website is apparently being added to, based on our observations accessing the website on a number of different occasions.

The languages we have included in our multilingual parallel corpus include those both ancient and modern, and are as follows:

| **Language** | **No. of translations** | **Used in tests** |
|---|---|---|
| Afrikaans | 1 | 12+ |
| Albanian | 1 | 27+ |
| Arabic | 1 | All |
| Chinese (Simplified) | 1 | 44+ |
| Chinese (Traditional) | 1 | 44+ |
| Croatian | 1 | 27+ |
| Czech | 2 | 12+ |
| Danish | 1 | 12+ |
| Dutch | 1 | 12+ |
| English | 7 | All |
| Finnish | 3 | 27+ |
| French | 2 | All |
| German | 4 | 8,27+ |
| Greek (New Testament) | 2 | 46+ |
| Hebrew (Old Testament) | 1 | 46+ |
| Hebrew (Modern) | 1 | 6,12+ |
| Hungarian | 1 | 6+ |
| Italian | 2 | 8,27+ |
| Japanese* | 1 | 9+ |
| Korean | 1 | 27+ |
| Latin | 1 | 8,9,28+ |
| Maori | 1 | 7,8,9,27+ |
| Norwegian | 1 | 27+ |
| Polish* | 1 | 27+ |
| Portuguese | 1 | 27+ |
| Russian | 1 | All |
| Spanish | 2 | All |
| Swedish | 1 | 27+ |
| Tagalog | 1 | 27+ |
| Thai | 1 | 27+ |
| Vietnamese | 1 | 27,44+ |

Table 1. Languages[1]

The languages above represent many of the major language groups: Austronesian (Maori and Tagalog); Altaic (Japanese and Korean); Sino-Tibetan (Chinese); Semitic (Arabic and Hebrew); Finno-Ugric (Finnish and Hungarian); Austro-Asiatic (Vietnamese); Tai-Kadai (Thai); and Indo-European (the remaining languages). The two New Testament Greek versions are the Byzantine/Majority Text (2000), and the parsed version of the same text, in which we treated distinct morphological elements (such as roots or inflectional endings) as distinct terms. Overall, the list includes

[1] Translations in languages marked with an asterisk above were obtained from websites other than the 'Unbound Bible' website. 'Used in tests' indicates in which tests in Table 2 below the language was used as training data, and hence the order of addition of languages to the training data.

47 versions in 31 distinct languages (assuming without further discussion here that each entry in the list represents a distinct language).

We aligned the translations by verse, and, since there are some differences in versification between translations (for example, the Hebrew Old Testament includes the headings for the Psalms as separate verses, unlike most translations), we spent some time cleaning the data to ensure the alignment was as good as possible, given available resources and our knowledge of the languages. (Even after this process, the alignment was not perfect, and differences in how well the various translations were aligned may account for some of the variability in the outcome of our experiments, depending on which translations were used.) The end result was that our parallel corpus consisted of 31,226 'mini-documents' – the total number of text chunks[2] after the cleaning process, aligned across all 47 versions. The two New Testament Greek versions, and the one Old Testament Hebrew version, were exceptions because these are only partially complete; the former have text in only 7,953 of the verses, and the latter has text in 23,266 of the verses. For some versions, a few of the verse translations are incomplete where a particular verse has been skipped in translation; this also explains the fact that the number of Hebrew and Greek text chunks together do not add up to 31,226. However, the number of such verses is negligible in comparison to the total.

## 3 Framework

The framework we used was the standard LSI framework described in Berry et al. (1994). Each aligned mini-document from the parallel corpus consists of the *combination* of text from all the 31 languages. A document-by-term matrix is formed in which each cell represents a weighted frequency of a particular term $t$ in a particular document $k$. We used a standard log-entropy weighting scheme, where the weighted frequency W is given by:

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

where F is the raw frequency of $t$ in $k$, $H_t$ is the standard '$p \log p$' measure of the entropy of the term across all documents, and N is the number of documents in the corpus. 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 sparse document-by-term 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. When we had a smaller number of languages in the mix, it was possible to use SVDPACK (Berry et al. 1996), which is an open-source non-parallel algorithm for computing the SVD, but for larger problems (involving more than a couple of dozen parallel versions), use of a parallel algorithm (in a library called Trilinos) was necessitated. (This was run on a Linux cluster consisting of 4,096 dual CPU compute nodes, running on Dell PowerEdge 1850 1U Servers with 6GB of RAM.)

In order to test the precision versus recall of our framework, we used translations of the 114 suras of the Qu'ran into five languages, Arabic, English, French, Russian and Spanish. The number of documents used for testing is fairly small, but large enough to give comparative results for our purposes which are still highly statistically significant. The test set was split into each of the 10 possible language-pair combinations: Arabic-English, Arabic-French, English-French, and so on.

For each language pair and test, 228 distinct 'queries' were submitted – each query consisting of one of the 228 sura 'documents'. If the highest-ranking document in the other language of the pair was in fact the query's translation, then the result was deemed 'correct'. To assess the aggregate performance of the framework, we used two measures: average precision at 0 (the maximum precision at any level of recall), and average precision at 1 document (1 if the 'correct' document ranked highest, zero otherwise). The second measure is a stricter one, but we generally found that there is a high rate of correlation between the two measures anyway.

## 4 Results and Discussion

The following tables show the results of our tests. First, we present in Table 2 the overall summary,

[2] The text chunks generally had the same boundaries as the verses in the original text.

with averages across all language pairs used in testing.

| No. of parallel versions | Average precision | |
|---|---|---|
| | At 0 | at 1 doc. |
| 2 | 0.706064 | 0.571491 |
| 3 | 0.747620 | 0.649269 |
| 4 | 0.617615 | 0.531873 |
| 5 | 0.744951 | 0.656140 |
| 6 | 0.811666 | 0.732602 |
| 7 | 0.827246 | 0.753070 |
| 8 | 0.824501 | 0.750000 |
| 9 | 0.823430 | 0.746053 |
| 12 | 0.827761 | 0.752632 |
| 27 | 0.825577 | 0.751316 |
| 28 | 0.823137 | 0.747807 |
| 44 | 0.839346 | 0.765789 |
| 46 | 0.839319 | 0.766667 |
| 47 | 0.842936 | 0.774561 |

Table 2. Summary results for all language pairs

From the above, the following should be clear: as more parallel translations are added to the index, the average precision rises considerably at first, and then begins to level off after about the seventh parallel translation. The results will of course vary according to which combination of translations is selected for the index. The number of such combinations is generally very large: for example, with 47 translations available, there are 47! / (40! 7!), or 62,891,499, possible ways of selecting 7 translations. Thus, for any particular number of parallel versions, we had to use some judgement in which parallel versions to select, since there was no way to achieve anything like exhaustive coverage of the possibilities.

Further, with more than 7 parallel translations, there is certainly no justification for saying that adding more translations or languages increases the 'noise' for languages in the test set, since beyond 7 the average precision remains fairly level. If anything, in fact, the precision still appears to rise slightly. For example, the average precision at 1 document rises by more than 0.75 percentage points between 46 and 47 versions. Given that in each of these experiments, we are measuring precision 228 times per language pair, and therefore 2,280 times in total, this small rise in precision is significant ($p \approx 0.034$). Interestingly, the 47$^{th}$ version to be added was parsed New Testament Greek. It appears, therefore, that the parsing helped in particular; we also have evidence from other experiments (not presented here) that overall precision is generally improved for *all* languages when Arabic wordforms are replaced by their respective citation forms (the bare root, or stem) – also a form of morphological parsing. Ancient Greek, like Arabic, is morphologically highly complex, so it would be understandable that parsing (or stemming) would help when parsing of either language is used in training.

One other point needs to be made here: the three versions added after the 44$^{th}$ version were the three incomplete versions (the two Greek versions cover just the New Testament, while Ancient Hebrew covers just the Old Testament). The above-mentioned increase in precision which resulted from the addition of these three versions is clear evidence that even in the case where a parallel corpus is defective for some language(s), including those languages can still result in the twofold benefit that (1) those languages are now available for analysis, and (2) precision is maintained or increased for the remaining languages.

Finally, precision at 1 document, the stricter of the two measures, is by definition less than or equal to precision at 0. This taken into account, it is also interesting that the gap between the two measures seems to narrow as more parallel translations and parsing are added, as Figure 1 shows.

For certain applications where it is important that the translation is ranked first, not just highly, among all retrieved documents, there is thus a particular benefit in using a 'massively parallel' aligned corpus.

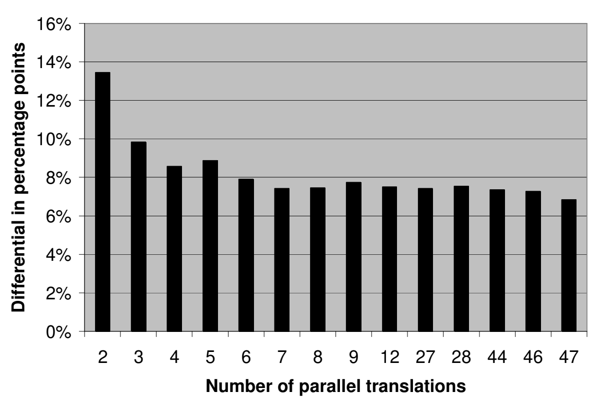


Figure 1. Differential between precision at 0 and precision at 1 document, by number of languages

Now we move on to look at more detailed results by language pair. Figure 2 below breaks down the results for precision at 1 document by language pair. In all tests, the two languages in each pair were (naturally) always included in the languages used for training. There is more volatility in the results by language pair than there is in the overall results, shown again at the right of the graph, which should come as no surprise since the averages are based on samples a tenth of the size. Generally, however, the pattern is the same for particular language pairs as it is overall; the more parallel versions are used in training, the better the average precision.

There are some more detailed observations which should also be made from Figure 2. First, the average precision clearly varies quite widely between language pairs. The language pairs with the best average precision are those in which two of English, French and Spanish are present. Of the five languages used for testing, these three cluster together genetically, since all three are Western (Germanic or Romance) Indo-European languages. Moreover, these are the three languages of the five which are written in the Roman alphabet.

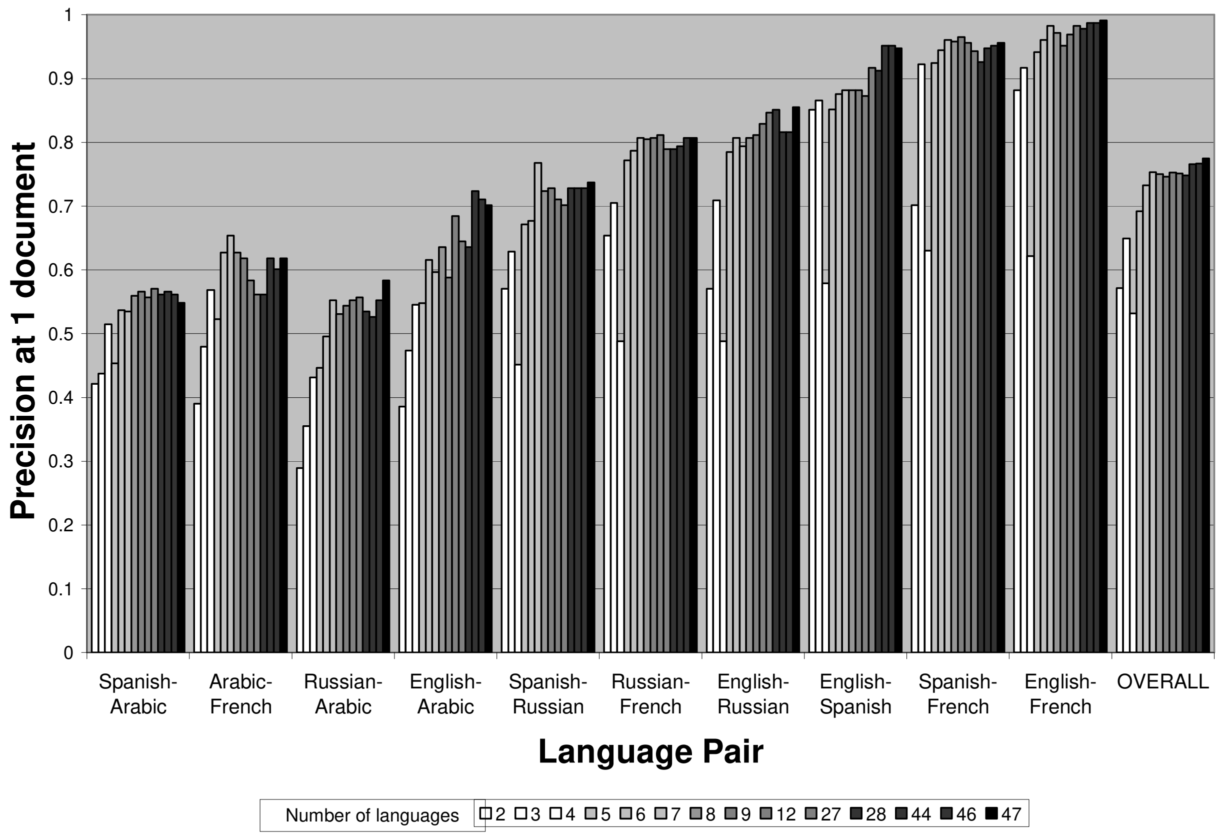


Figure 2. Chart of precision at 1 doc. by language pair and number of parallel training versions

However, we believe the explanation for the poorer results for language pairs involving either Arabic, Russian, or both, can be pinned down to something more specific. We have already partially alluded to the obvious difference between Arabic and Russian on the one hand, and English, French and Spanish on the other: that Arabic and Russian are highly morphologically rich, while English, French and Spanish are generally analytic languages. This has a clear effect on the statistics for the languages in question, as can be seen in Table 3, which is based on selected translations of the Bible for each of the languages in question.

| Translation | Types | Tokens |
|---|---|---|
| English (King James) | 12,335 | 789,744 |
| Spanish (Reina Valera 1909) | 28,456 | 704,004 |
| Russian (Synodal 1876) | 47,226 | 560,524 |
| Arabic (Smith Van Dyke) | 55,300 | 440,435 |
| French (Darby) | 20,428 | 812,947 |

Table 3. Statistics for Bible translations in 5 languages used in test data

Assuming that the respective translations are faithful (and we have no reason to believe otherwise), and based on the statistics in Table 3, it should be the case that Arabic contains the most 'information' per term (in the information theoretic sense), followed by Russian, Spanish, English and French.[3] Again, this corresponds to our intuition that much information is contained in Arabic patterns and Russian inflectional morphemes, which in English, French and Spanish would be contained in separate terms (for example, prepositions).

Without additional pre-processing, however, LSI cannot deal adequately with root-pattern or inflectional morphology. Moreover, it is clearly a weakness of LSI, or at least the standard log-entropy weighting scheme as applied within this framework, that it makes no adjustment for differences in information content per word between languages. Even though we can assume near-equivalency of information content between the different translations above, according to the standard log-entropy weighting scheme there are large differences between the total entropy of particular parallel documents; in general, languages such as English are overweighted while those such as Arabic are underweighted.

Now that this issue is in perspective, we should draw attention to another detail in Figure 2. Note that the language pairs which benefited most from the addition of Ancient Greek and Hebrew into the training data were those which included Russian, and Russian-Arabic saw the greatest increase in precision. Recall also that the 47th version to be added was the parsed Greek, so that essentially each Greek morpheme is represented by a distinct term. From Figure 2, it seems clear that the inclusion of parsed Greek in particular boosted the precision for Russian (this is most visible at the right-hand side of the set of columns for Russian-Arabic and English-Russian). There are, after all, notable similarities between modern Russian and Ancient Greek morphology (for example, the nominal case system). Essentially, the parsed Greek acts as a 'clue' to LSI in associating inflected forms in Russian with preposition/non-inflected combinations in other languages. These results seem to be further confirmation of the notion that parsing just one of the languages in the mix helps overall; the greatest boost is for those languages with morphology related to that of the parsed language, but there is at least a maintenance, and perhaps a small boost, in the precision for unrelated languages too.

Finally, we turn to look at some effects of the particular languages selected for training. Included in the results above, there were three separate tests run in which there were 6 training versions. In all three, Arabic, English, French, Russian and Spanish were included. The only factor we varied in the three tests was the sixth version. In the three tests, we used Modern Hebrew (a Semitic language, along with Arabic), Hungarian (a Uralic language, not closely related to any of the other five languages), and a second English version respectively. The results of these tests are shown in Figure 3, with figures for the test in which only 5 versions were included for comparative purposes.

From these results, it is apparent first of all that it was generally beneficial to add a sixth version, regardless of whether the version added was English, Hebrew or Hungarian. This is consistent with the results reported elsewhere in this paper. Second, it is also apparent that the greatest benefit overall was had by using an additional English version, rather than using Hebrew or Hungarian. Moreover, perhaps surprisingly, the use of Hebrew in training – even though Hebrew is related to Arabic – was of less benefit to Arabic than either Hungarian or an additional English version. It appears that the use of multiple versions in the same language is beneficial because it enables LSI to make use of the many different instantiations in the expression of a concept in a single language, and that this effect can be greater than the effect which obtains from using heterogeneous languages, even if there is a genetic relationship to existing languages.

[3] To clarify the meaning of 'term' here: for all languages except Chinese, text is tokenized in our framework into terms using regular expressions; each non-word character (such as punctuation or white space) is assumed to mark the boundary of a word. For Chinese, we made the simplifying assumption that each character represented a separate term.

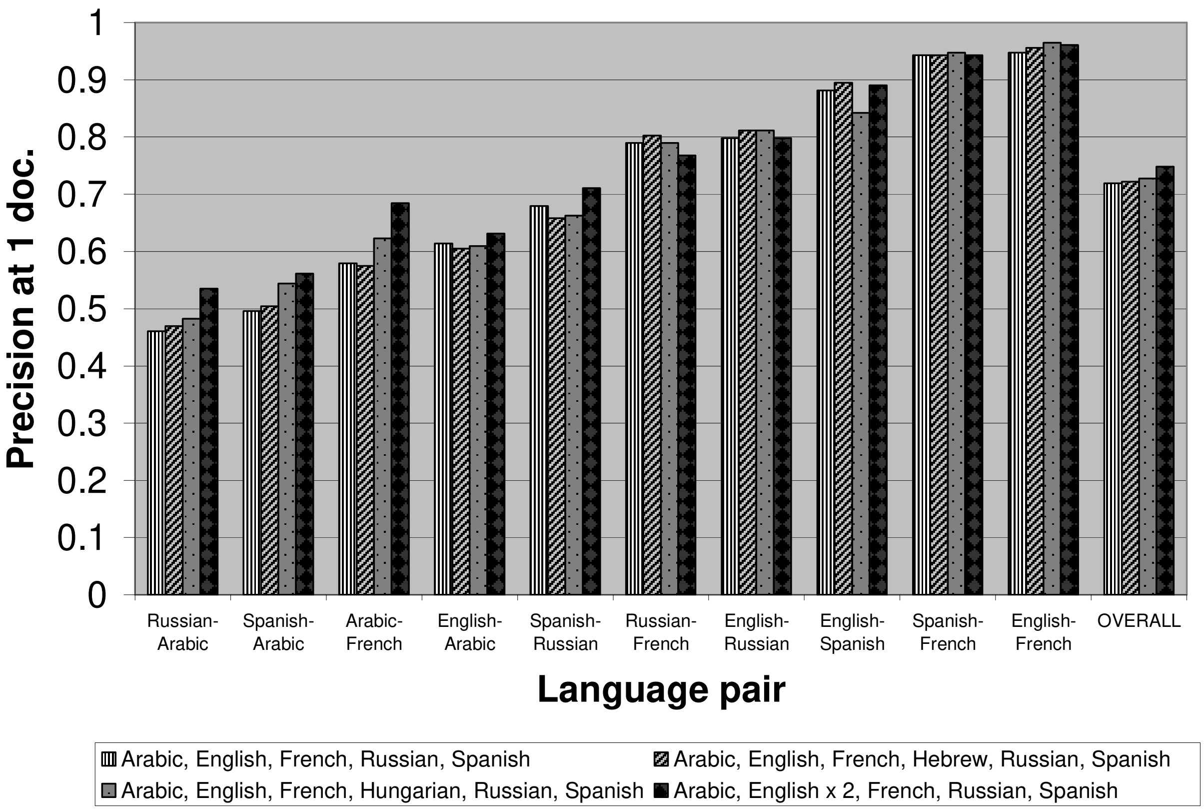


Figure 3. Precision at 1 document for 6 training versions, with results of using different mixes of languages for training

Figure 3 may also shed some additional light on one other detail from Figure 2: a perceptible jump in precision between 28 and 44 training versions for Arabic-English and Arabic-French. It should be mentioned that among the 16 additional versions were five English versions (American Standard Version, Basic English Bible, Darby, Webster's Bible, and Young's Literal Translation), and one French version (Louis Segond 1910). It seems that Figure 2 and Figure 3 both point to the same thing: that the use of parallel versions or translations in a *single* language can be particularly beneficial to overall precision within the LSI framework – even to a greater extent than the use of parallel translations in different languages.

## 5 Conclusion

In this paper, we have shown how 'massive parallelism' in an aligned corpus can be used to improve the results of cross-language information retrieval. Apart from the obvious advantage (the ability to automate the processing of a greater variety of linguistic data within a single framework), we have shown that including more parallel translations in training improves the precision of CLIR across the board. This is true whether the additional translations are in the language of another translation already within the training set, whether they are in a related language, or whether they are in an unrelated language; although this is not to say that these choices do not lead to (generally minor) variations in the results. The improvement in precision also appears to hold whether the additional translations are complete or incomplete, and it appears that morphological pre-processing helps, not just for the languages pre-processed, but again across the board.

Our work also offers further evidence that the supply of useful pre-existing parallel corpora is not perhaps as scarce as it is sometimes claimed to be. Compilation of the 47-version parallel corpus we used was not very time-consuming, especially if the time taken to clean the data is not taken into account, and all the textual material we used is publicly available on the World Wide Web.

While the experiments we performed were on non-standard test collections (primarily because the Qu'ran was easy to obtain in multiple languages), it seems that there is no reason to believe our general observation – that more parallelism in the training data is beneficial for cross-language retrieval – would not hold for text from other domains. Whether the genre of text used as training data affects the *absolute* rate of retrieval precision for text of a different genre (e.g. news articles, shopping websites) is a separate question, and one we intend to address more fully in future work.

In summary, it appears that we are able to achieve the results we do partly because of the inherent properties of LSI. In essence, when the data from more and more parallel translations are subjected to SVD, the LSI 'concepts' become more and more reinforced. The resulting trend for precision to increase, despite 'blips' for individual languages, can be seen for all languages. To put it in more prosaic terms, the more different ways the same things are said in, the more understandable they become – including in cross-language information retrieval.

## 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.

## References

Lars Asker. 2004. Building Resources: Experiences from Amharic Cross Language Information Retrieval. Paper presented at *Cross-Language Information Retrieval and Evaluation: Workshop of the Cross-Language Evaluation Forum, CLEF 2004*.

Ricardo Baeza-Yates and Berthier Ribeiro-Neto. 1999. *Modern Information Retrieval*. New York: ACM Press.

Michael Berry, Theresa Do, Gavin O'Brien, Vijay Krishna, and Sowmimi Varadhan. 1996. SVDPACKC (Version 1.0) User's Guide. Knoxville, TN: University of Tennessee.

Bible Society. 2006. *A Statistical Summary of Languages with the Scriptures*. Accessed at http://www.biblesociety.org/latestnews/latest341-slr2005stats.html on Jan. 5, 2007.

Biola University. 2005-2006. *The Unbound Bible*. Accessed at http://www.unboundbible.com/ on Jan. 5, 2007.

Peter Chew, Stephen Verzi, Travis Bauer and Jonathan McClain. 2006. Evaluation of the Bible as a Resource for Cross-Language Information Retrieval. In *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*, 68-74. Sydney: Association for Computational Linguistics.

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

Julio Gonzalo. 2001. Language Resources in Cross-Language Text Retrieval: a CLEF Perspective. In Carol Peters (ed.). *Cross-Language Information Retrieval and Evaluation: Workshop of the Cross-Language Evaluation Forum, CLEF 2000*: 36-47. Berlin: Springer-Verlag.

Dragos Munteanu and Daniel Marcu. 2006. Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. *Computational Linguistics* 31(4):477-504.

Jian-Yun Nie and Fuman Jin. 2002. A Multilingual Approach to Multilingual Information Retrieval. *Proceedings of the Cross-Language Evaluation Forum*, 101-110. Berlin: Springer-Verlag.

Jian-Yun Nie, Michel Simard, Pierre Isabelle, and Richard Durand. 1999. Cross-Language Retrieval based on Parallel Texts and Automatic Mining of Parallel Texts from the Web. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 74-81, August 15-19, 1999, Berkeley, CA.

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

Recherche appliquée en linguistique informatique (RALI). 2006. *Corpus aligné bilingue anglais-français*. Accessed at http://rali.iro.umontreal.ca/ on February 22, 2006.

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