

Evaluation of the Bible as a Resource for Cross-Language Information Retrieval

**Association for Computational Linguistics
Multilingual Resources and Interoperability Workshop
July 23, 2006**

**Peter Chew, Steve Verzi, JT McClain, Travis Bauer
Sandia National Laboratories**

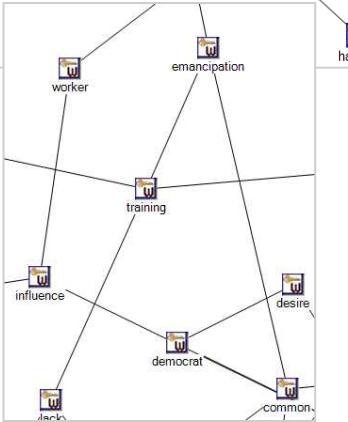
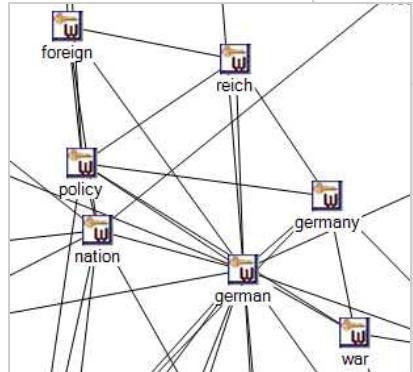


Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company,
for the United States Department of Energy under contract DE-AC04-94AL85000.

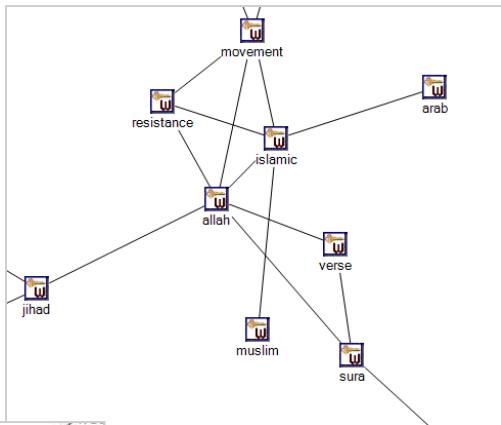




Background – our use for CLIR



words

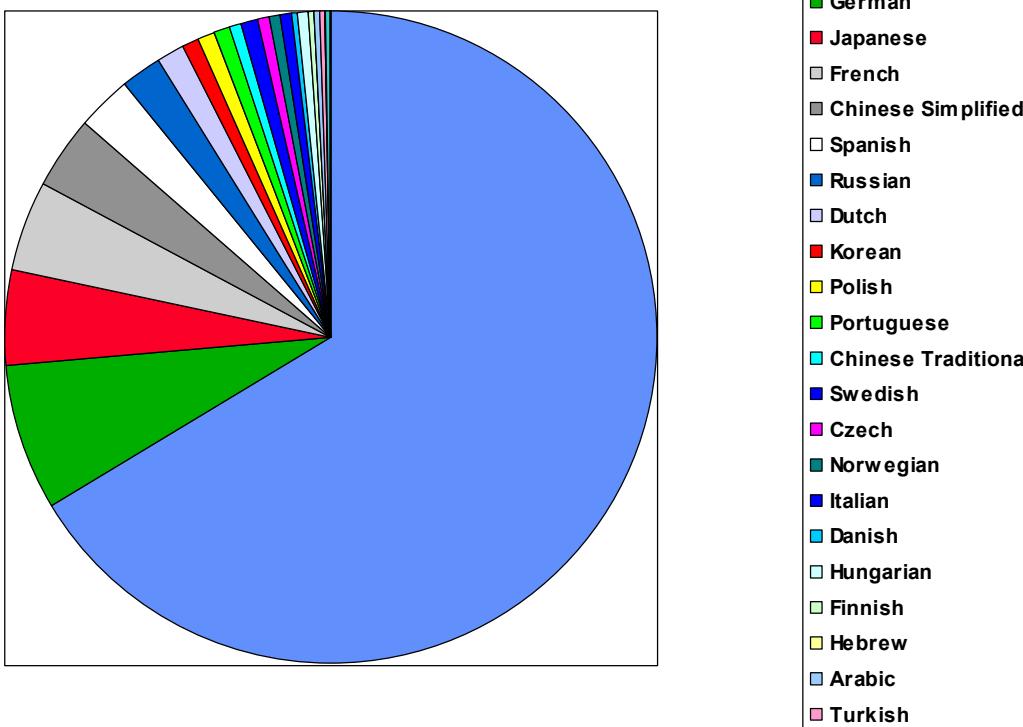


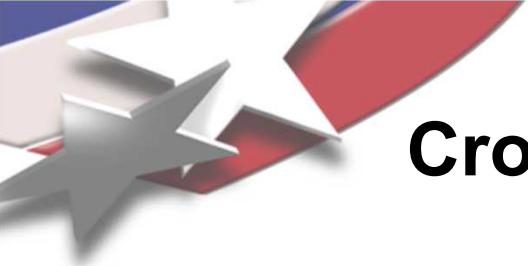
ideas



movements

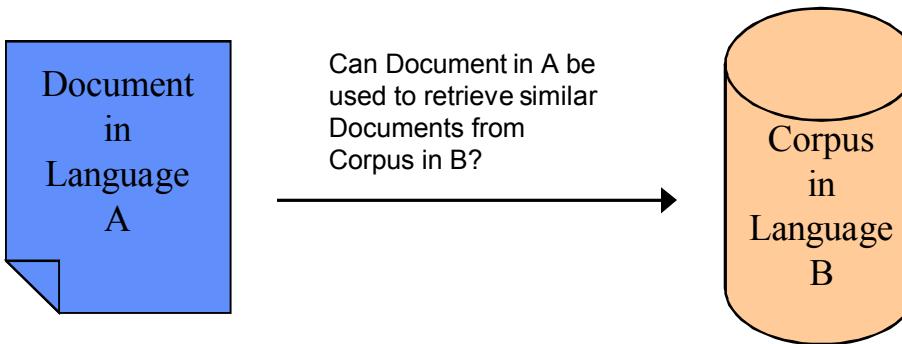
Distribution of internet content by language



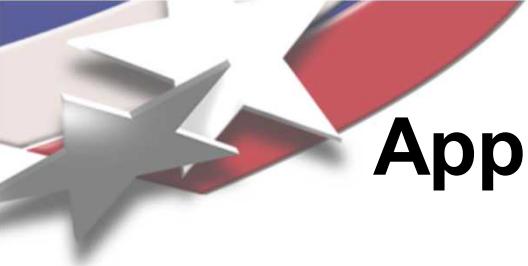


Cross-Language Comparison – Experimental Design

- **Basic Question**
 - Given 2 documents in separate languages, can we determine if they are “about” the same thing?



- **Basic Test**
 - Given a document in language A, compute similarity between it and each document in a corpus of language B

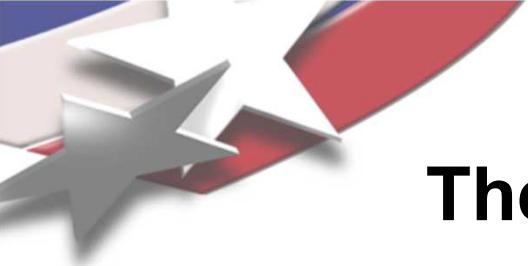


Approaches to cross-language information retrieval

- Translate the query
 - Efficacy is constrained by quality of machine translation
- Train algorithm on parallel corpora
 - Translations should:
 - Be available in target languages
 - Be reliable
 - Be sufficiently large in size
 - Cover target subject domain
 - Be free of undue copyright restrictions
 - Be electronically available
 - Be alignable



```
health of animals act ===> loi
health of the economy ===> santé
heavy burden ===> lourd fardeau
heavy equipment ===> matériel lourd
heritage day ===> jour du patrimoine
higher premiums ===> cotisation
highest bidder ===> plus offrant
historic document ===> document
historic event ===> événement
house management committee ===>
house of commons standing commit
```



The Bible as a Parallel Corpus

- Resnik, Olsen & Diab (1999) showed that the Bible fulfills all of these criteria and is surprisingly suitable as a parallel corpus
 - Translations in > 2,400 languages and rising
 - Great care taken over translations
 - Respectably large compared to other corpora
 - Covers many modern genres
 - Covers up to 85% of modern vocabulary
 - Generally free of copyright restrictions
 - Electronically available
 - Alignable



Language coverage - detail

A STATISTICAL SUMMARY OF LANGUAGES WITH THE SCRIPTURES

A summary, by geographical area and type of publication, of the number of different languages and dialects in which publication of at least one book of the Bible had been registered as of December 31, 2005.

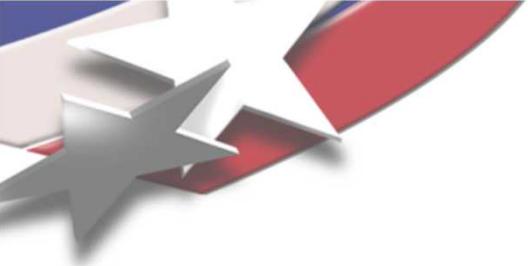
| Continent/Region | Portions | Testaments | Bibles | Bibles, DC* | Total |
|--|------------|--------------|------------|----------------|--------------|
| Africa | 223 | 301 | 159 | (29) | 683 |
| Asia | 218 | 244 | 131 | (28) | 593 |
| Australia/New Zealand/ Pacific Islands | 148 | 234 | 38 | (9) | 420 |
| Europe | 114 | 36 | 61 | (47) | 211 |
| North America | 39 | 30 | 7 | (0) | 76 |
| Caribbean Islands / Central America / Mexico/South America | 118 | 270 | 29 | (9) | 417 |
| Constructed Languages | 2 | 0 | 1 | (0) | 3 |
| TOTALS | 862 | 1,115 | 426 | (122) | 2,403 |

* This column is a sub-section of the Bibles column – for example, there is a translation of the Deuterocanon for 47 of the 61 languages of Europe in which the Bible has been translated.

[A few corrections were made to our language databases and are reflected in this statistical summary]

Per <http://www.biblesociety.org/latestnews/latest341-slr2005stats.html>

The 'Unbound Bible'

A graphic of the American flag, featuring stars and stripes, is positioned in the top left corner of the slide.

The Unbound Bible

File Edit View Favorites Tools Help

Address http://www.unboundbible.com/index.cfm?method=downloads.showDownloadMain

[English] [Русский] [한국어]
[Translate] [Nederlands] [Français] [עברית] [English]

Bible Search Bible Study Tools Know God Downloads Translate
Bible in a Year Search Settings FAQ Links Contact
Download Unicode Fonts and Bibles

Unicode Bibles

Afrikaans 1953 † Download

Afrikaans 1953 †
Albanian †
Amharic NT
Arabic: Smith & Van Dyke †
Aramaic NT: Peshitta †
Armenian (Eastern): (Genesis, Exodus, Gospels)
Armenian (Western): NT (Dwight/Riggs, 1853)
Basque (Navarro-Labourdin): NT
Breton Gospels
Chamorro (Psalms, Gospels, Acts)
Chinese: NCV (Traditional) †
Chinese: Union (Simplified) †
Chinese: NCV (Simplified) †
Chinese: Union (Traditional) †
Croatian 2.0
Czech BKR †
Czech CEP †
Czech KMS †
Czech NKB †
Danish †
Dutch Staten Vertaling †
English: King James Version 2.0
English: American Standard Version
English: Basic English Bible
English: Darby Version
English: Douay-Rheims 2.0
English: Webster's Bible
English: Weymouth NT
English: World English Bible
English: Young's Literal Translation

mir Rybant has written a free software program that works with Bibles in the Unbound Bible format. You can look up a passage, and then copy and paste it into your favorite word processor. The program is available for download below.

UniBible - A Multilingual Bible Reader for PalmOS

UniBible is a multilingual Bible-reader program for the PalmOS platform. It uses Unicode to display Bible texts in various languages, including:

Armenian, Arabic, Chinese, English, French, German, Hebrew, Italian, Japanese, Korean, Portuguese, Russian, Spanish, and more. The program is available for download for free.

Unicode Fonts

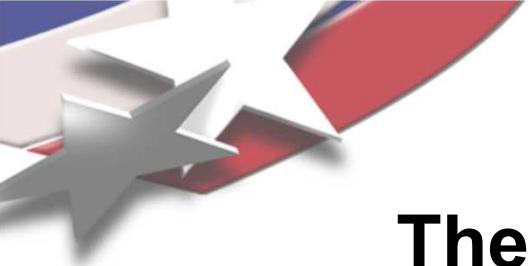
UniBible uses Unicode fonts to display the Bible text. The fonts are available for download for free.

GB18030 Support Package

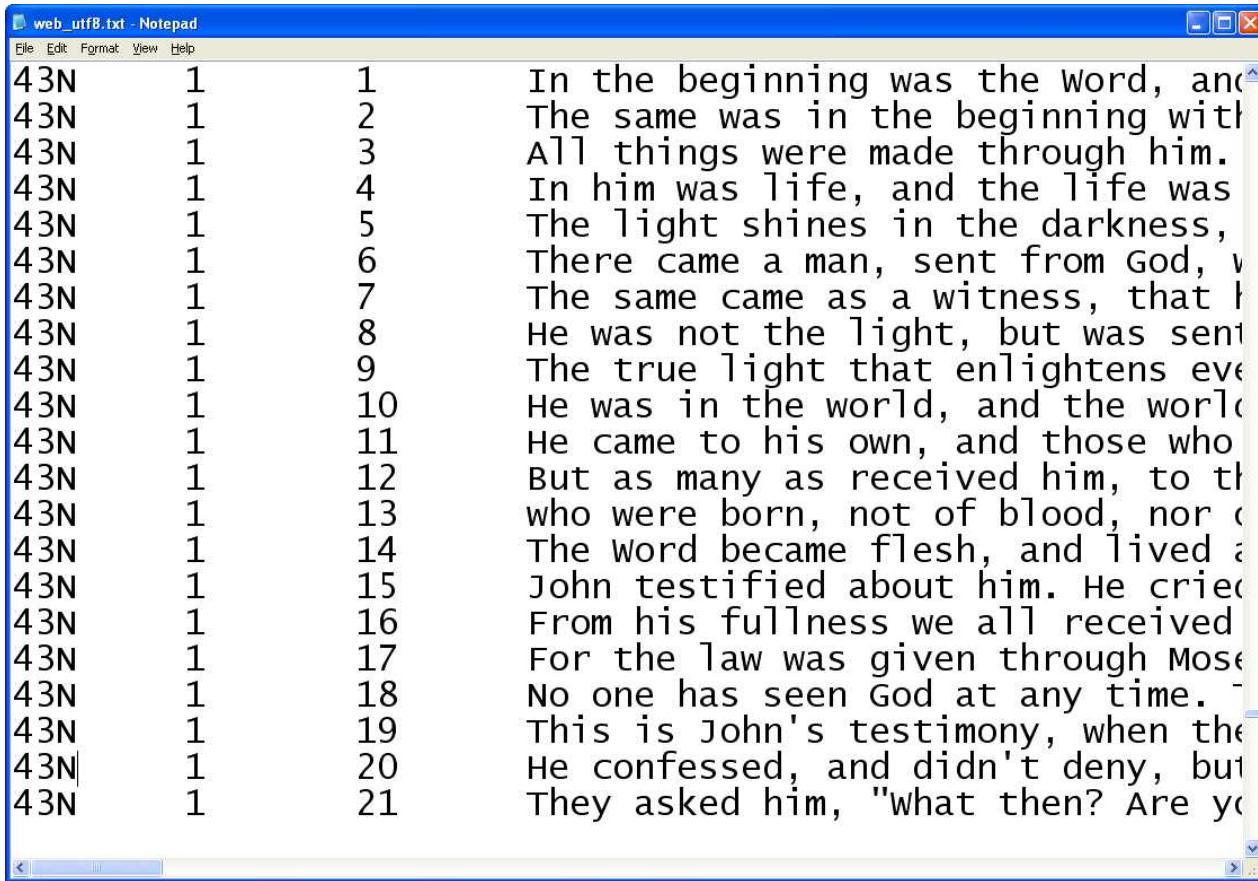
UniBible supports the GB18030 character encoding, which is used in China. The support package is available for download for free.

**85 translations (some partial) in
51 languages, in common format**

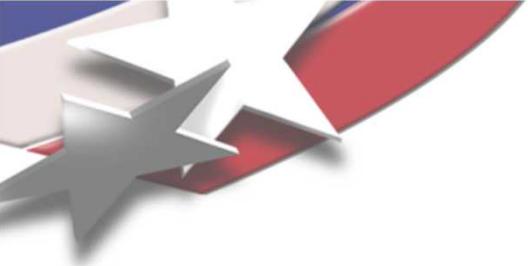
Internet



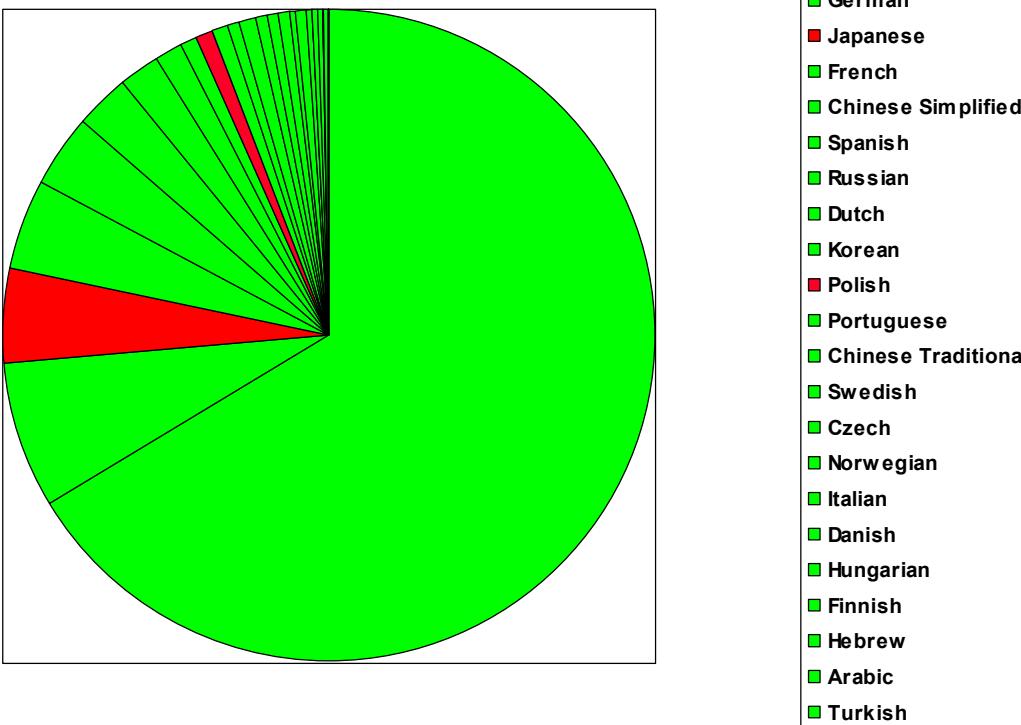
The 'Unbound Bible' – a sample

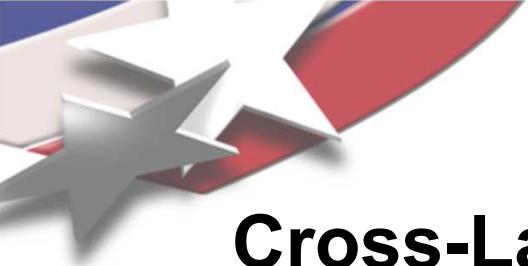


43N 1 1 In the beginning was the Word, and the Word was with God, and the Word was God.
43N 1 2 The same was in the beginning with God, and the Word was God.
43N 1 3 All things were made through him, and without him was not any thing made that was made.
43N 1 4 In him was life, and the life was the light of men.
43N 1 5 The light shines in the darkness, and the darkness comprehended it not.
43N 1 6 There came a man, sent from God, whose name was John.
43N 1 7 The same came as a witness, testifying about the light, that all might believe through him.
43N 1 8 He was not the light, but was sent to testify about the light.
43N 1 9 The true light that enlightens everyone was coming into the world.
43N 1 10 He was in the world, and the world did not know him.
43N 1 11 He came to his own, and those who were his own did not receive him.
43N 1 12 But as many as received him, to them he gave the right to become children of God, even to those who were born, not of blood, nor of the will of man, nor of the will of God.
43N 1 13 The Word became flesh, and lived among us, full of grace and truth.
43N 1 14 John testified about him. He cried out, saying, "This is the Lamb of God who takes away the sin of the world."
43N 1 15 From his fullness we have all received, grace upon grace.
43N 1 16 For the law was given through Moses; the promise of the grace of God came through the Christ.
43N 1 17 No one has seen God at any time. If you have seen me, you have seen God; and he is in me and I am in him.
43N 1 18 This is John's testimony, when the Jews sent priests and Levites from Jerusalem to ask him, "Who are you?"
43N 1 19 He confessed, and didn't deny, but said, "I am not the Christ."
43N 1 20 They asked him, "What then? Are you Elijah?" He said, "I am not." They said, "Are you the prophet?" He said, "I am not."



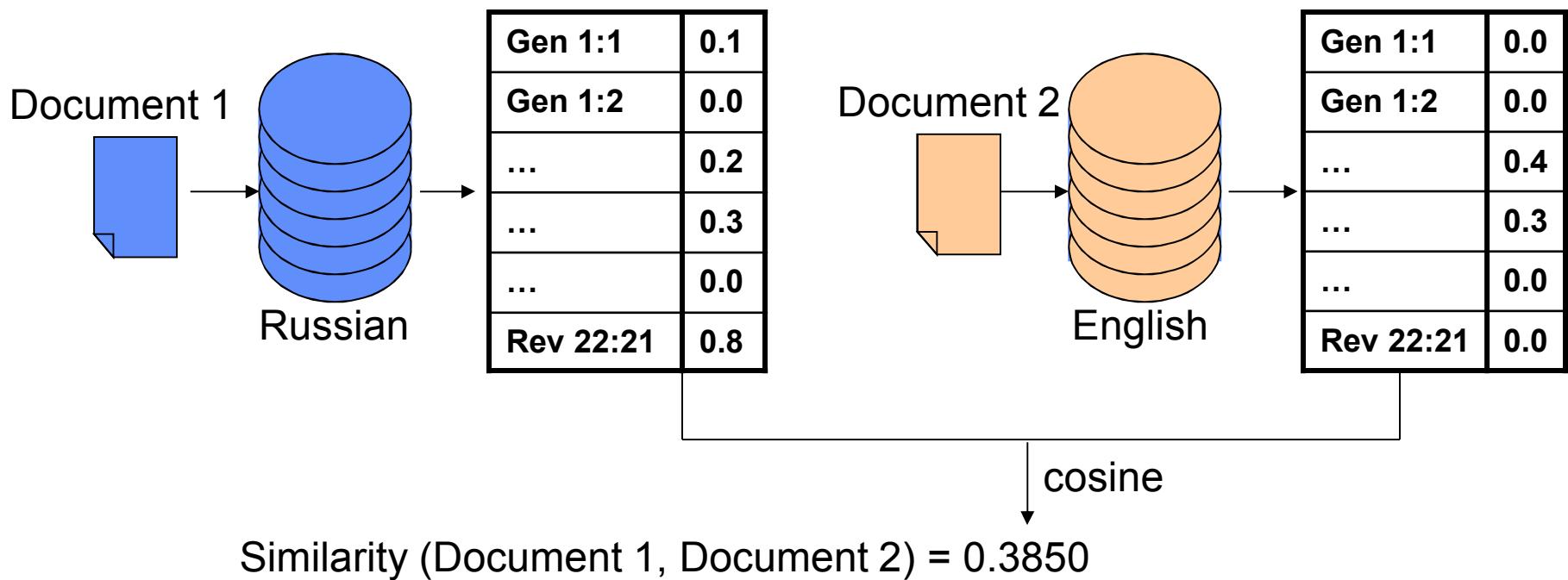
Coverage of internet content based on 'Unbound Bible'

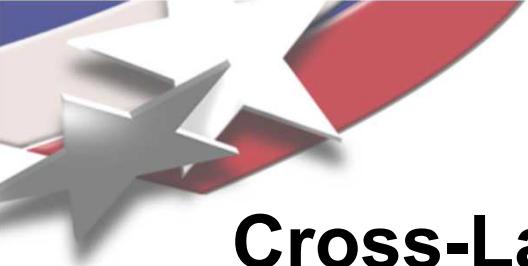




Cross-Language Comparison – Method 1

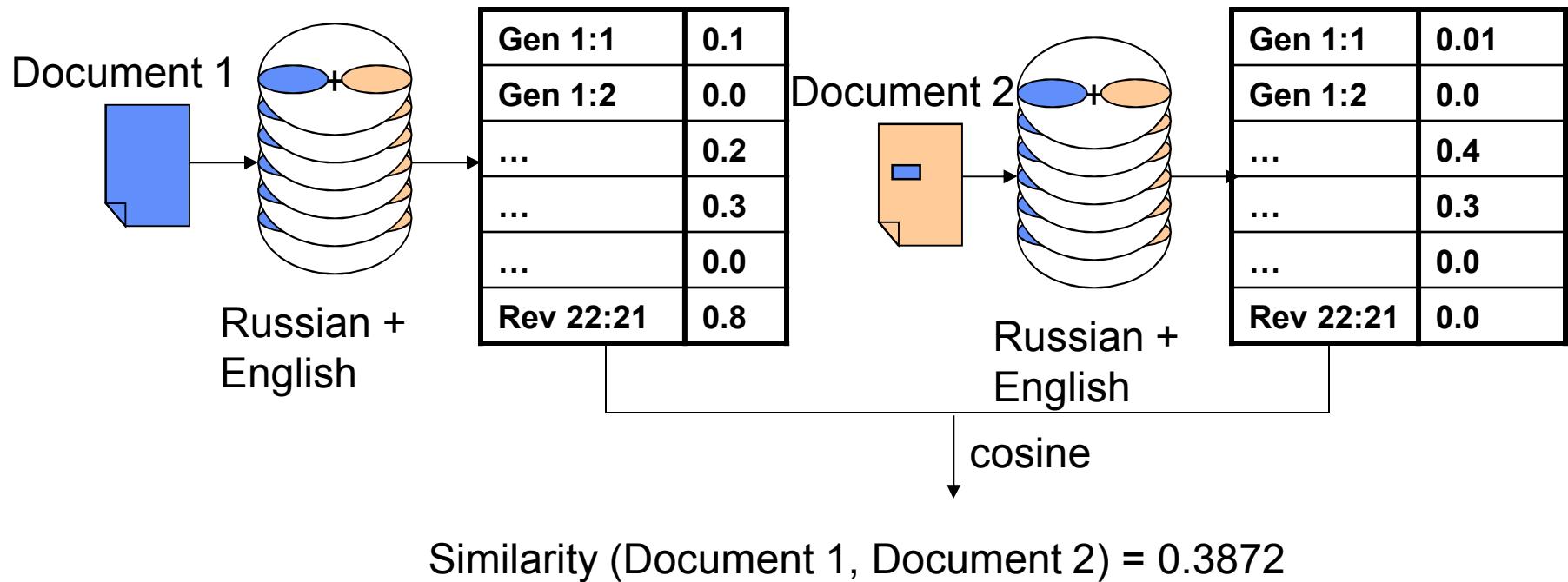
- Use separate index for each language





Cross-Language Comparison – Method 2

- Use single index for all languages (concatenated)





Validation results: Test set in training set

Method: Index on entire Bible, measure average uninterpolated precision at doc. 1 for 66 books of Bible

| | | To | | | | |
|------|---------|--------|---------|--------|---------|---------|
| | | Arabic | English | French | Russian | Spanish |
| From | Arabic | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | English | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | French | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | Russian | 1.00 | .99 | 1.00 | 1.00 | 1.00 |
| | Spanish | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

Validation results: Test set **not** in training set

Method: Index on entire Bible, obtain matrix of similarity measures for 5 conference abstracts where English and Spanish translations exist

| | | English | | | | |
|---------|-------|---------|-------|-------|-------|-------|
| | | Doc 1 | Doc 2 | Doc 3 | Doc 4 | Doc 5 |
| Spanish | Doc 1 | .607 | .043 | .045 | .035 | .022 |
| | Doc 2 | .049 | .397 | .038 | .082 | .166 |
| | Doc 3 | .030 | .050 | .045 | .101 | .049 |
| | Doc 4 | .102 | .096 | .080 | .189 | .105 |
| | Doc 5 | .035 | .131 | .039 | .042 | .168 |

We have also used the framework successfully for Maori, to distinguish between the Treaty of Waitangi and the New Zealand National Anthem

Validation results: Test set **not** in training set

Method: Index on entire Bible, measure Mean Average Precision for 114 suras of Quran in English, Arabic, Russian, and Spanish (results comparable to McNamee & Mayfield 2004)

| - LSA - no removal of stopwords | | To | | | |
|---------------------------------------|---------|---------|---------|---------|---------|
| | | Arabic | English | Russian | Spanish |
| From | Arabic | 1.00 | .71 .60 | .62 .33 | .72 .46 |
| | English | .71 .49 | 1.00 | .90 .75 | .90 .53 |
| | Russian | .56 .40 | .92 .68 | 1.00 | .67 .45 |
| | Spanish | .66 .46 | .87 .78 | .74 .62 | 1.00 |

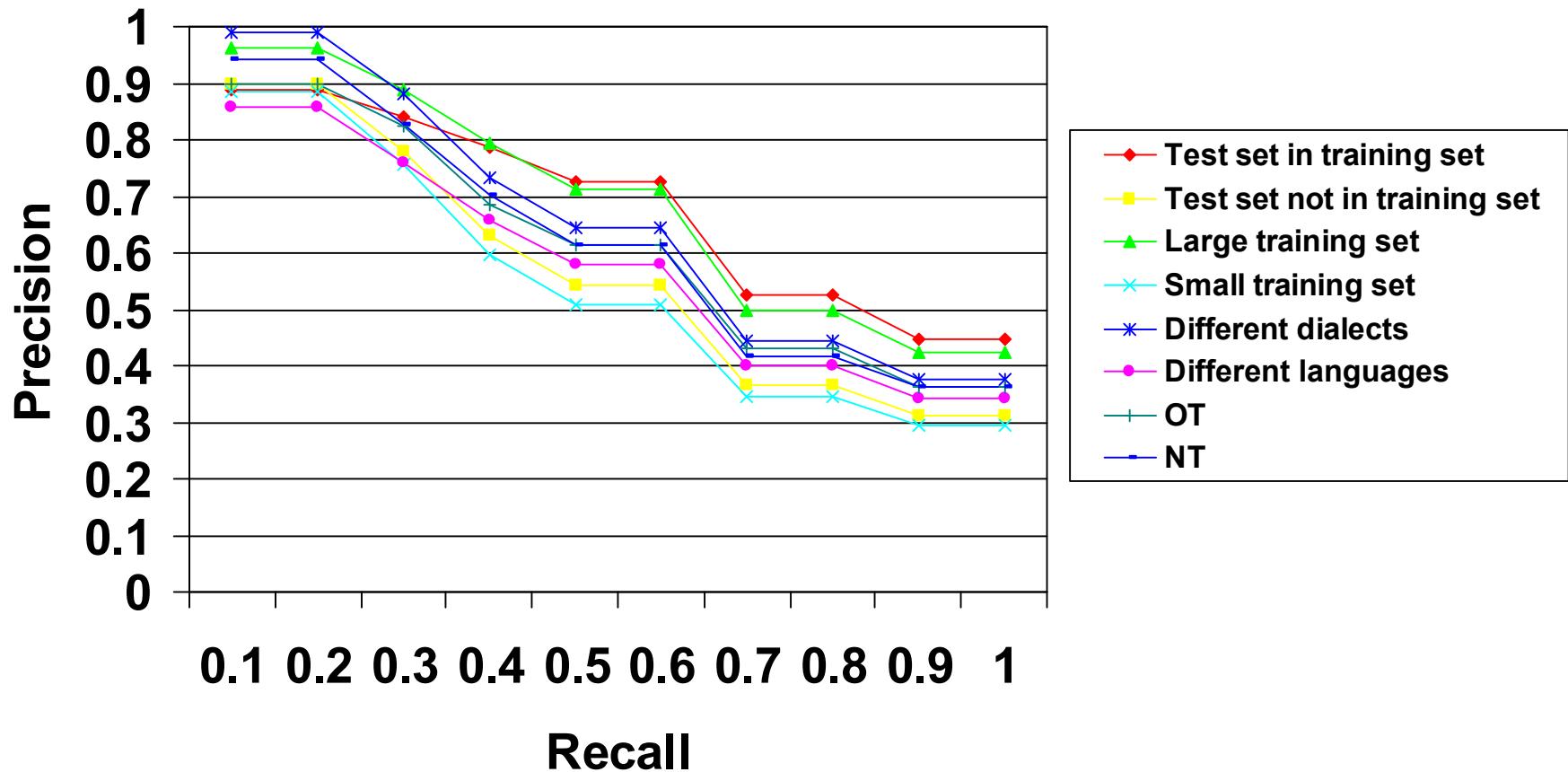
.35

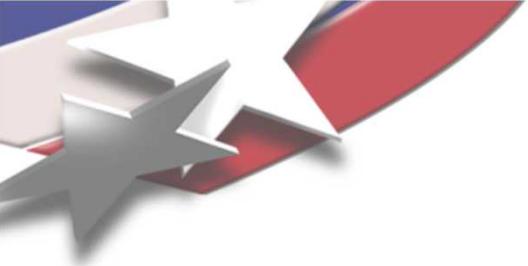
(using 5-
grams)

Method 1: Separate index for each language

Method 2: Single index for all languages

Precision-recall graphs for different test parameters





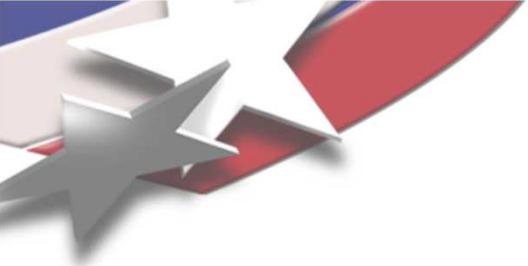
Observations

- Cross-language approach is easily extensible to new languages and corpora, including minority languages
- Resources exist which allow large parallel corpora to be built up from scratch in hours, at no monetary cost
- Unsurprisingly, the larger the training set, the better the precision-recall results
- Results appear to be comparable to, or better than, those achieved by other methods and reported recently in research literature; further testing may be needed
- Our best results were obtained using LSA, a single index for all languages, and without removing stopwords. This has the advantage of requiring no language-specific expertise to set up.



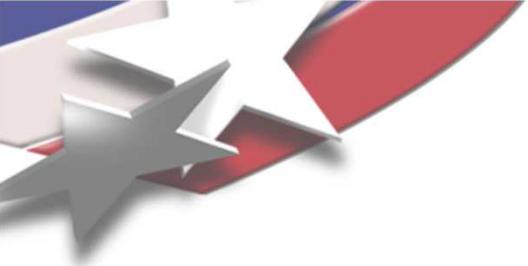
Future directions

- Continue to test CLIR algorithm to identify its strengths and weaknesses
- What chunk size yields best cross-language IR results, and why?
- Can we use the output of cross-language comparison to characterize documents by their ideology?



Selected References

- Ackland, R. 2005. Mapping the U.S. Political Blogosphere: Are Conservative Bloggers More Prominent? mimeo., The Australian National University.
- Chomsky, Noam. 1988 *Language and Politics*. C.P. Otero (ed.). Montreal: Black Rose Books.
- Landauer, Thomas. 1998. An Introduction to Latent Semantic Analysis. In *Discourse Processes* 25, 259-284.
- McNamee, Paul, and James Mayfield. 2004. Character N-Gram Tokenization for European Language Text Retrieval. *Information Retrieval* 7:73-97.
- Resnik, Philip, Mari Olsen, and Mona Diab. 1999. The Bible as a Parallel Corpus: Annotating the Book of 2000 Tongues. In *Computers and the Humanities* 33, 1-2. pp 129-153.
- Resnik, Philip, and Noah Smith. 2003. The Web as a Parallel Corpus. In *Computational Linguistics* 29, No. 3, pp. 349-380.



QUESTIONS?