SAND2006-4189C

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

Association for Computational Linguistics

Multilingual Resources and Interoperability Workshop

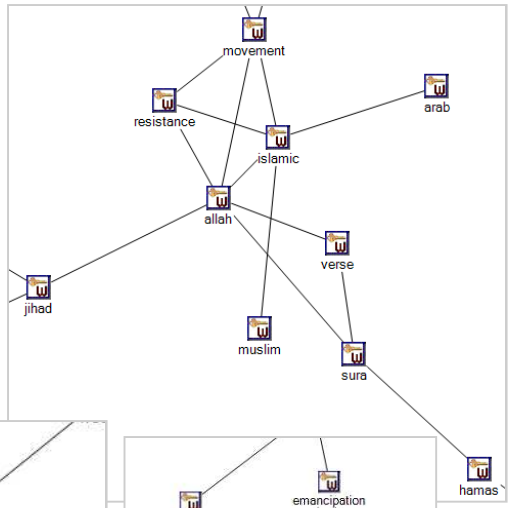
July 23, 2006

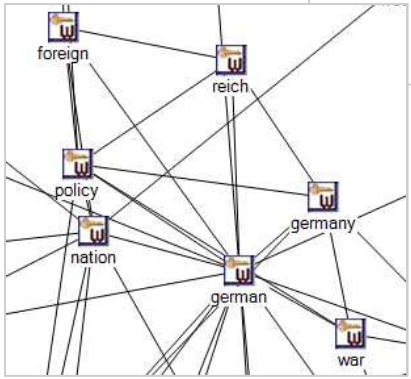
Peter Chew, Steve Verzi, JT McClain, Travis Bauer
Sandia National Laboratories

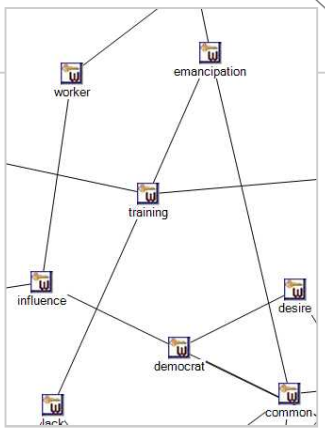
Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract DE-AC04-94AL85000.

Sandia National Laboratories

Background – our use for CLIR







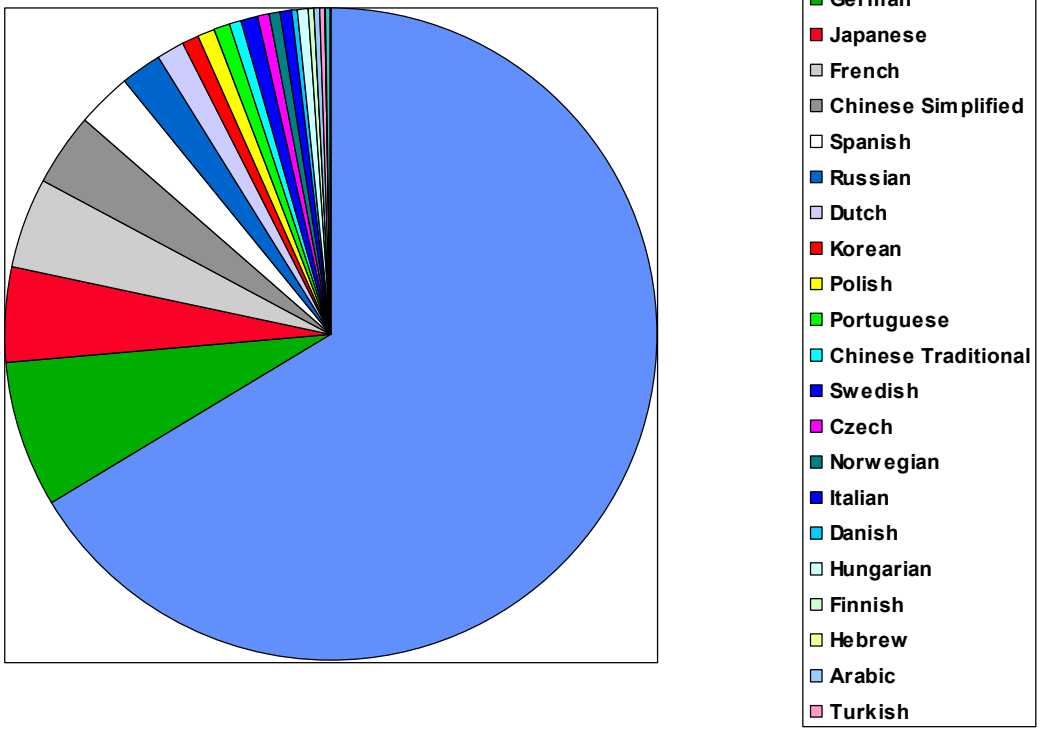
words

ideas

movements

Sandia National Laboratories

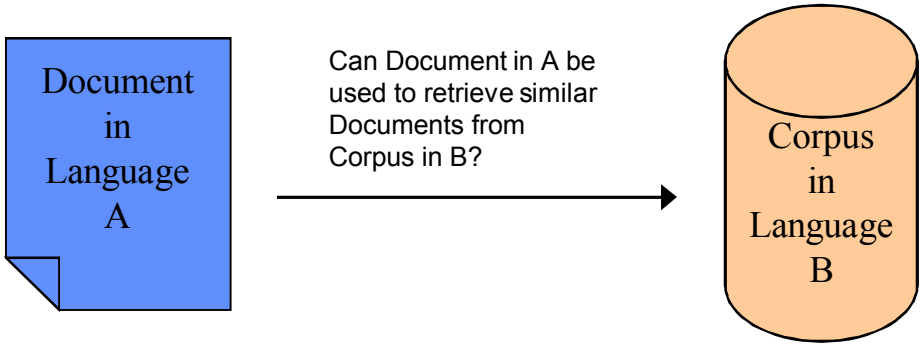
Distribution of internet content by language



Sandia National Laboratories

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  ===> sant
heavy burden  ===> lourd fardeau
heavy equipment  ===> matériel l
heritage day  ===> jour du patri
higher premiums  ===> cotisation
highest bidder  ===> plus offran
historic document  ===> document
historic event  ===> événement h
house management committee  ===>
house of commons standing commit
```

Sandia National Laboratories

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**

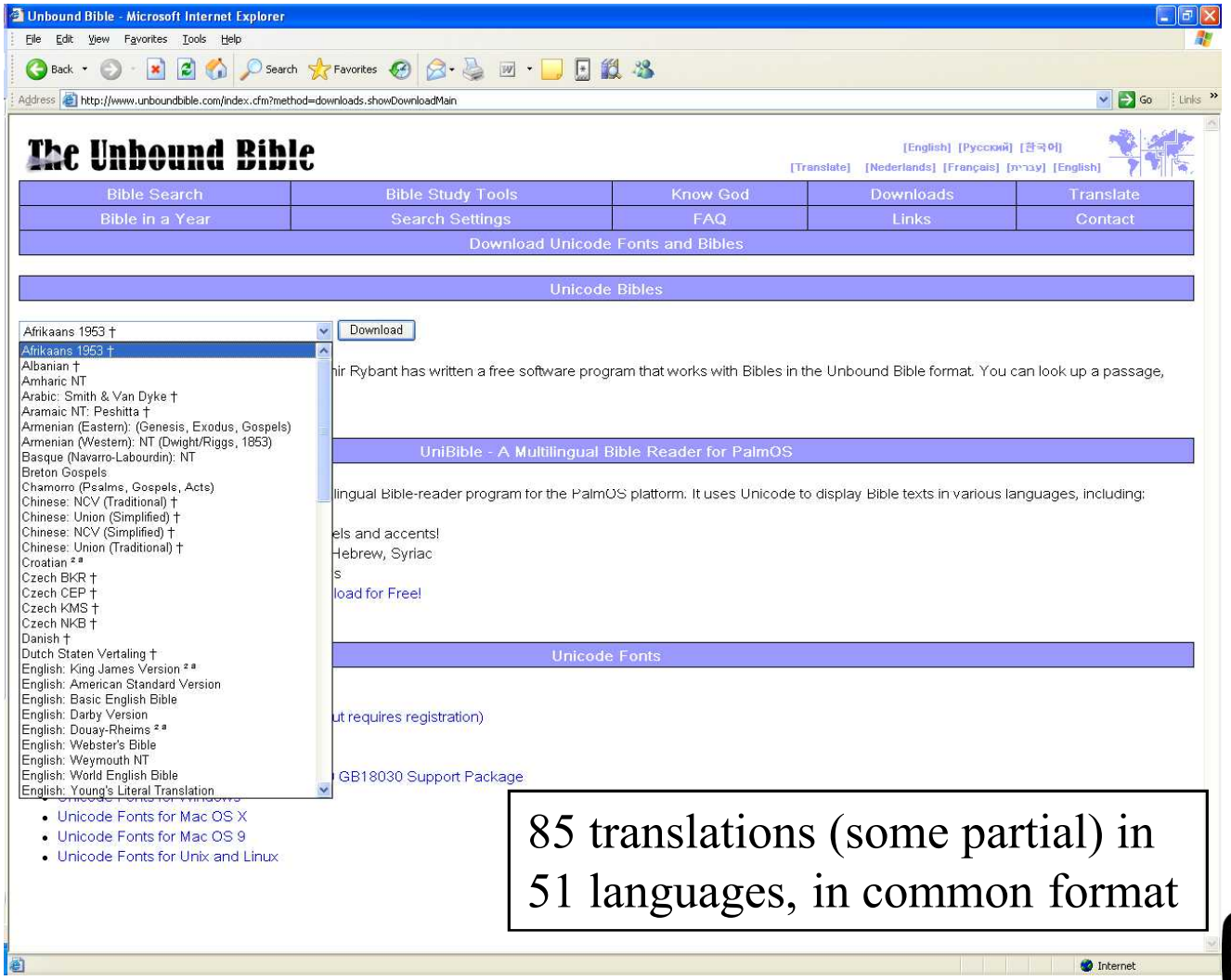
* This column is a sub-section of the Bibles column – for example, there is a translation of the Deuterocanon for 47 of the 61 languages of Europe in which the Bible has been translated.

[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

Sandia National Laboratories

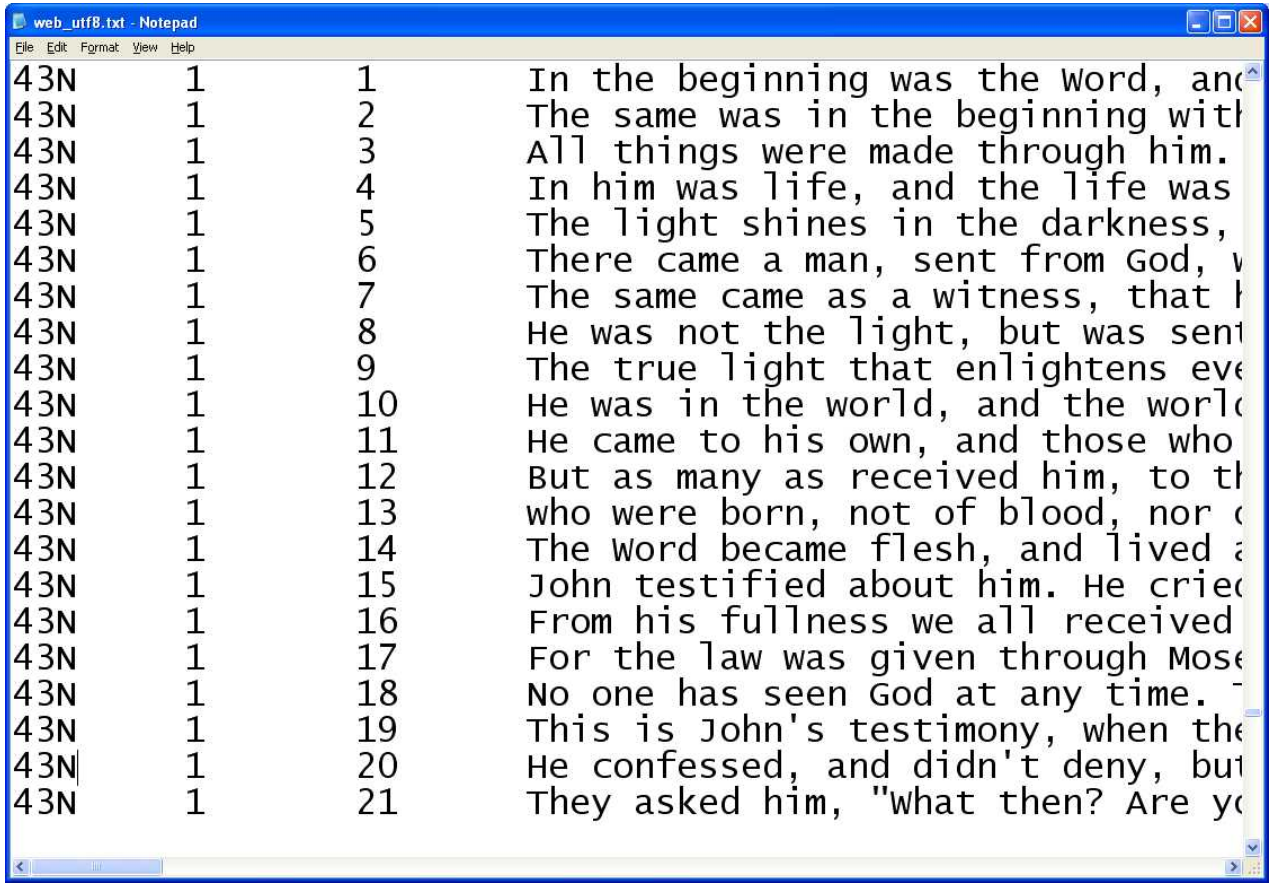
The ‘Unbound Bible’



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

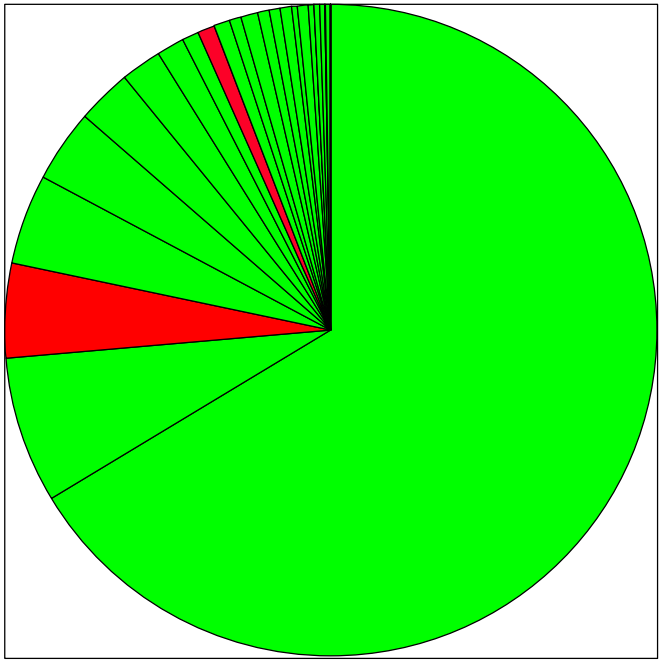
Sandia National Laboratories

The ‘Unbound Bible’ – a sample



Sandia National Laboratories

Coverage of internet content based on ‘Unbound Bible’

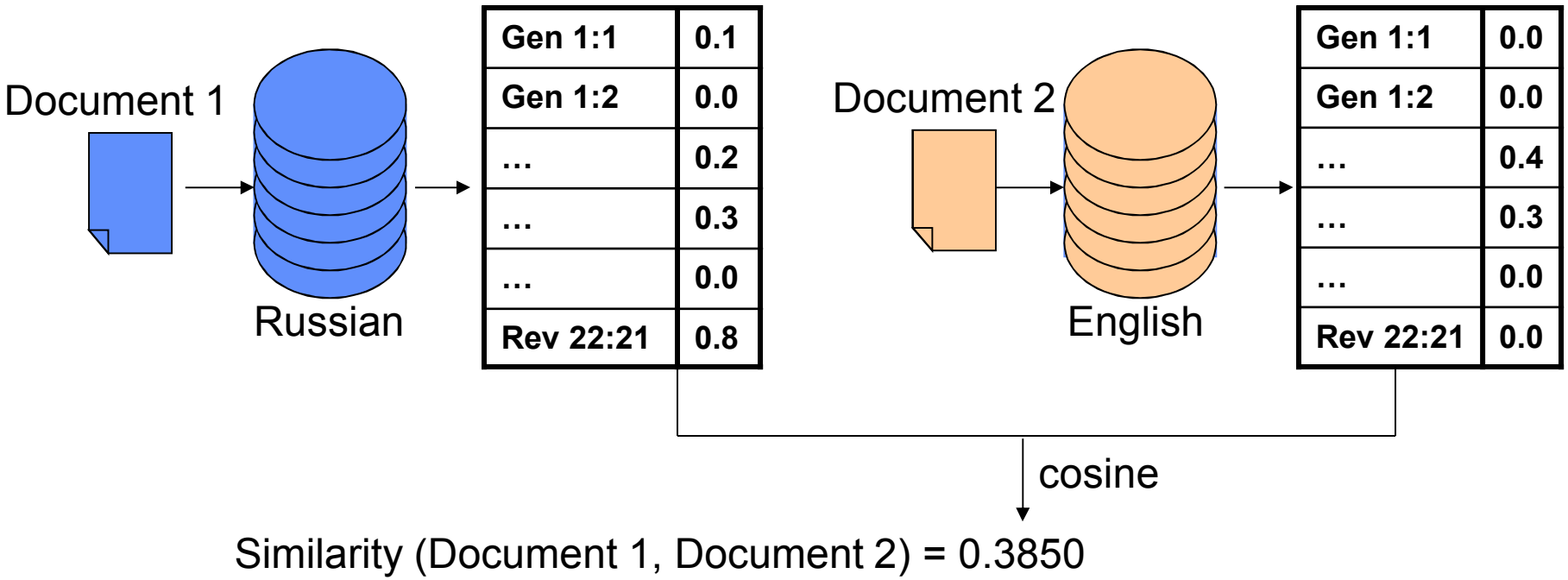


- English
- German
- Japanese
- French
- Chinese Simplified
- Spanish
- Russian
- Dutch
- Korean
- Polish
- Portuguese
- Chinese Traditional
- Swedish
- Czech
- Norwegian
- Italian
- Danish
- Hungarian
- Finnish
- Hebrew
- Arabic
- Turkish

Sandia National Laboratories

Cross-Language Comparison – Method 1

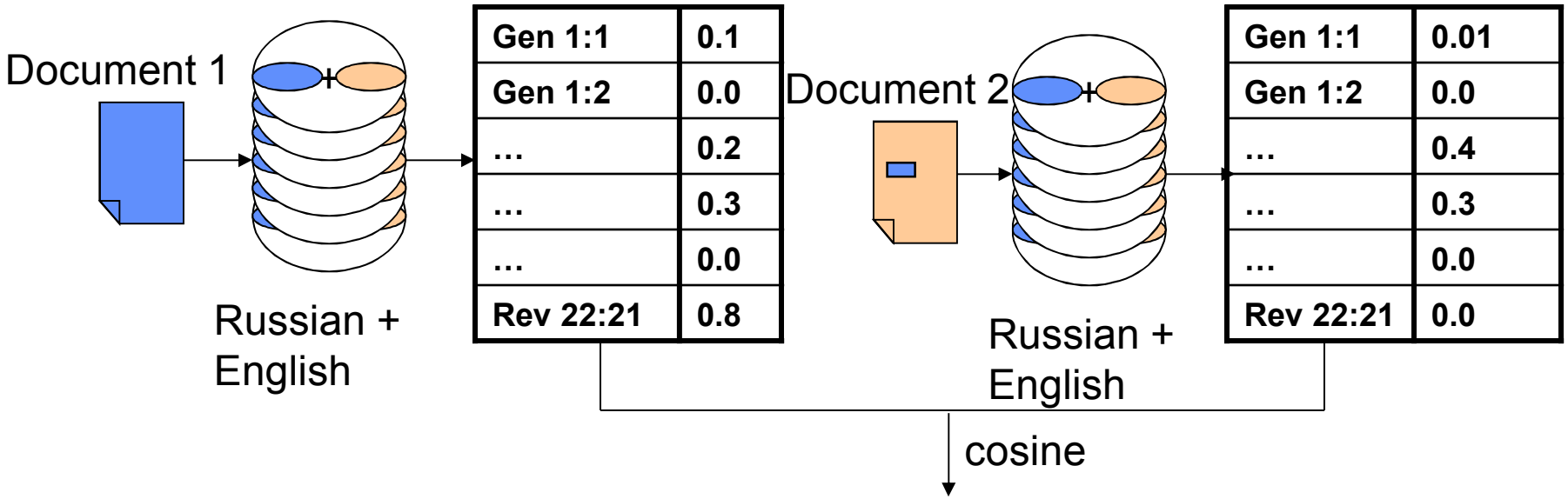
- **Use separate index for each language**



Sandia National Laboratories

Cross-Language Comparison – Method 2

- **Use single index for all languages (concatenated)**



Similarity (Document 1, Document 2) = 0.3872

Sandia National Laboratories

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

Sandia National Laboratories

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

Sandia National Laboratories

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

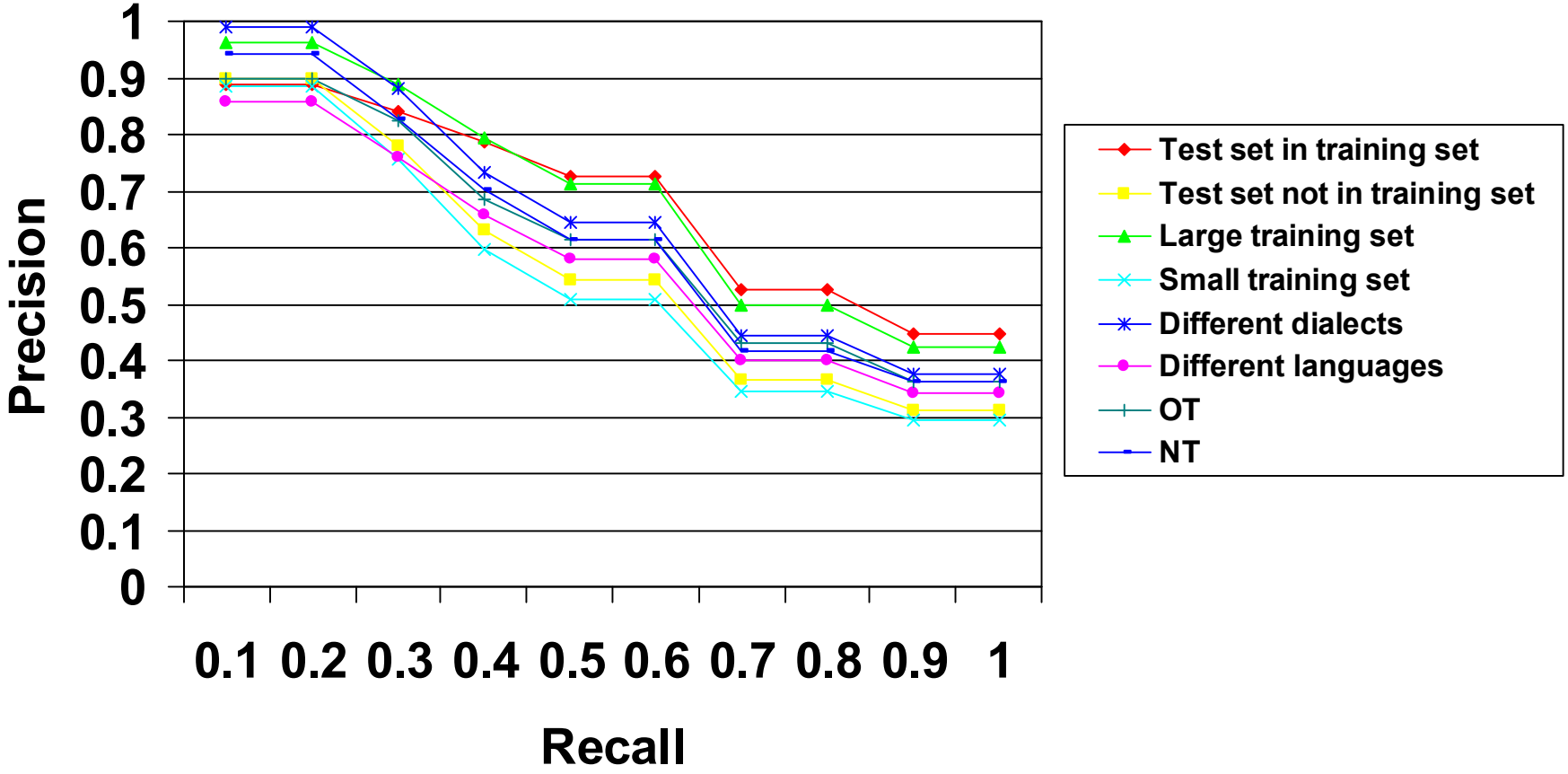
(using 5-grams)

Method 1: Separate index for each language

Method 2: Single index for all languages

Sandia National Laboratories

Precision-recall graphs for different test parameters



Sandia National Laboratories

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?

Sandia National Laboratories

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?

Sandia National Laboratories