SAND2007-3082P

High-Level Visualization of Multilingual Text from the Internet: Using Linguistic Methods to Identify Topics of Interest

Seminar at the
Physical Science Laboratory, NMSU

May 22, 2007

Peter Chew
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

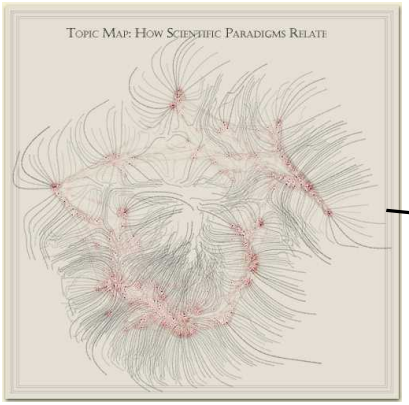
Vision

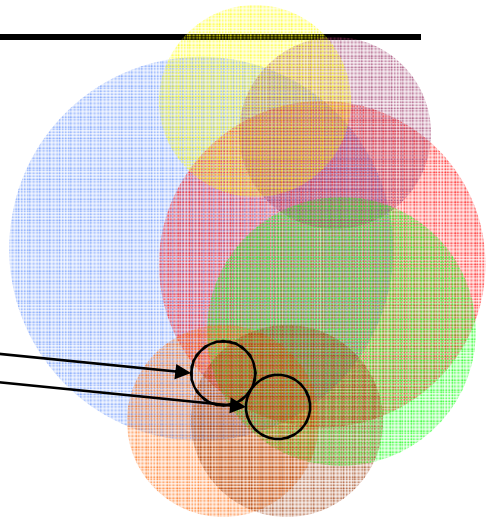
Imagine that you can create a representation, using a graphical or spatial layout, of all blogs and similar pages on the internet.

- What will it look like?
- What will it contain?
- How will it be used?
- What will people be able to learn from it?
- Can it be predictive?

- How do we make it?

Recent work at Sandia





ISI database

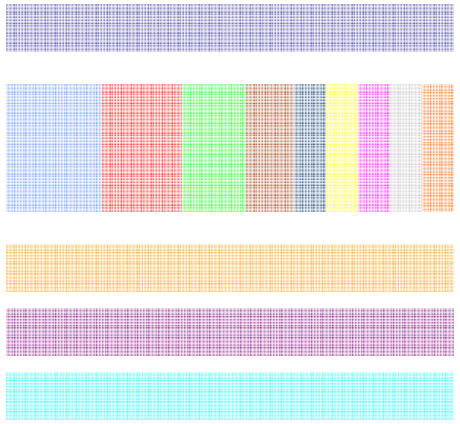
One year – 1M papers

Relatively clean, fielded data

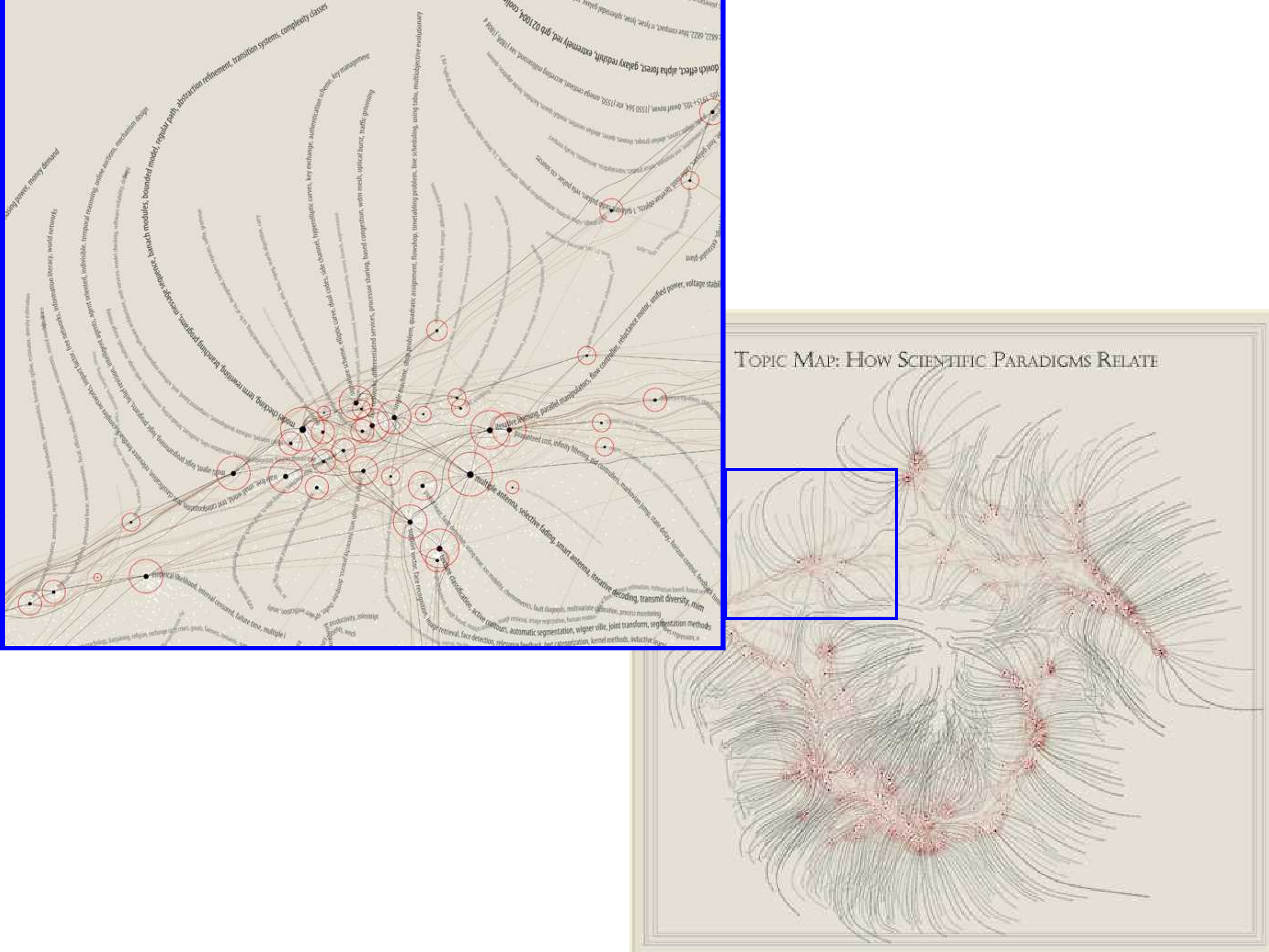
Matching refs to previous papers

Million node layout

- 100k clusters, very precise topic focus



Sandia National Laboratories

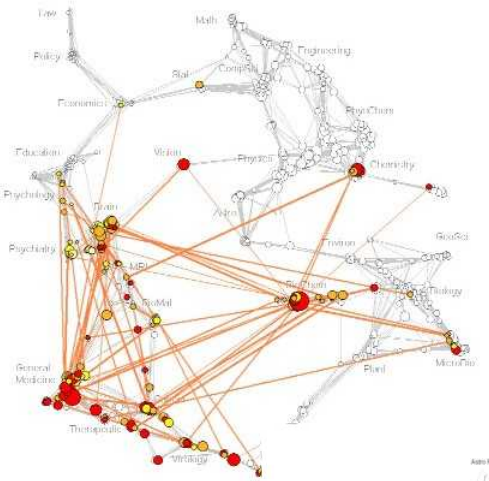


Topic Map: How Scientific Paradigms Relate
multiple antenna, selective fading, smart antenna, iterative decoding, transmit diversity, mim
message sequence, banach modules, bounded model, regular path, abstraction refinement, transition systems, complexity classes
model checking, term rewriting, branching programs
iterative learning, parallel manipulator, flow controller, reluctance motor, unified power, voltage stabi

Uses

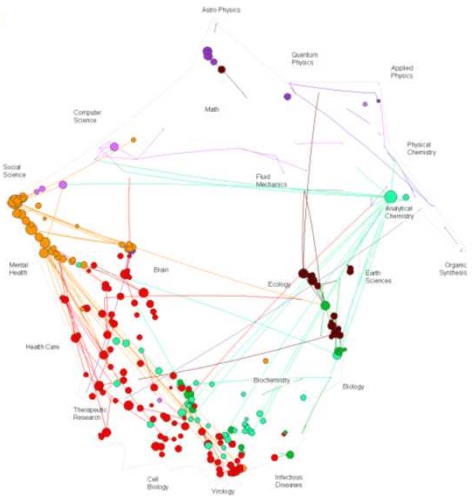
CURRENT

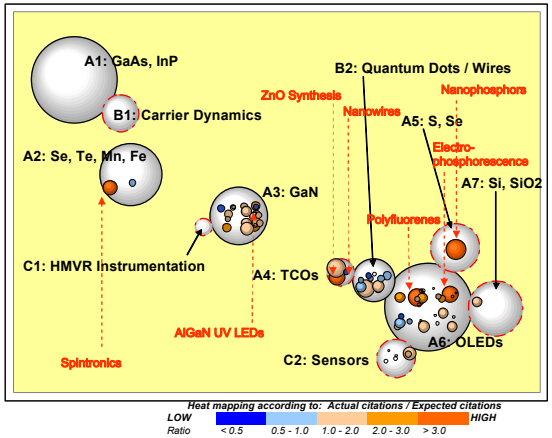
- Agency, corporate profiles
- National strength networks
- Opportunities for collaboration

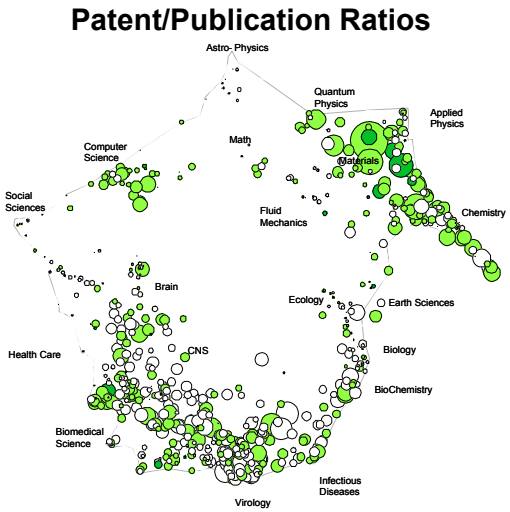


FUTURE?

- Early detection of significant socio-political shifts
- Tracking changes over time in public interest in certain topics

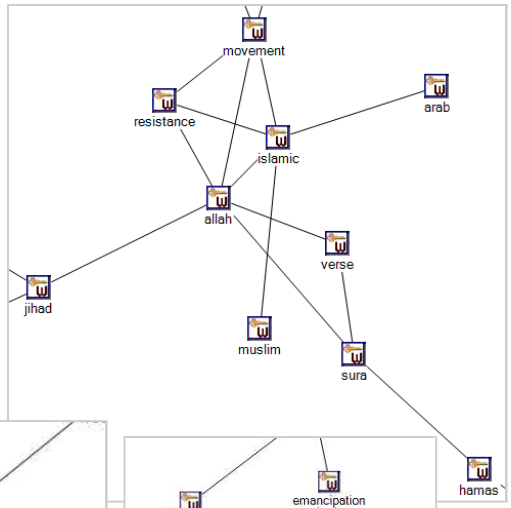


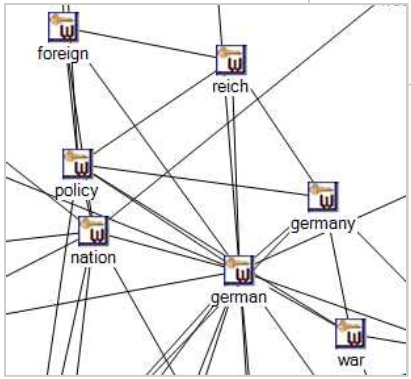


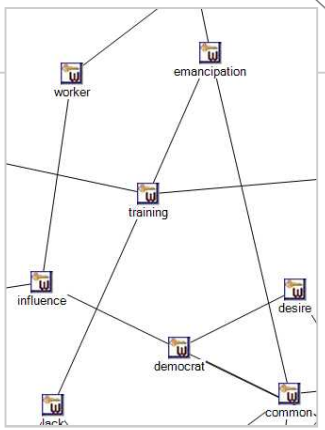


Sandia National Laboratories

The internet as a marketplace of ideas







words

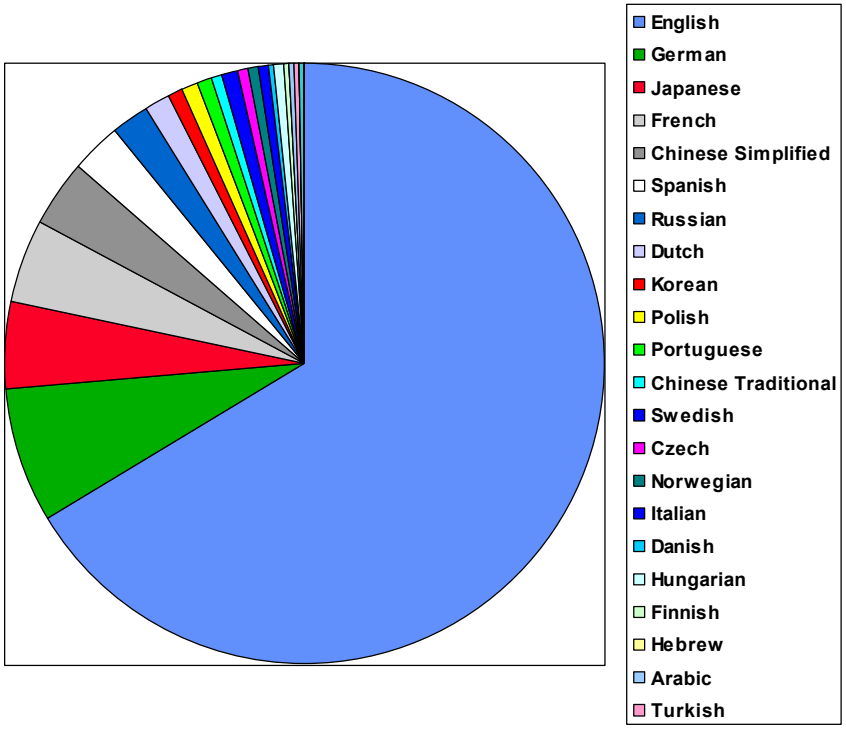
ideas

movements

Sandia National Laboratories

Multilingual information retrieval
(identification of similar documents in different languages)

Languages of the internet



Our framework:

- Identifies similar documents across language boundaries
- Can handle 81 languages
- Covers about 99.75% of internet content (based on chart at left)
- Is easily extensible to new languages
- Identifies translations of documents with up to 99% precision
- Enables users to find documents of interest in an unknown language

Sandia National Laboratories

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 'Rosetta Stone' (1)

- 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

Sandia National Laboratories

The Bible as a ‘Rosetta Stone’ (2)

Language	Text	Words	%
EN	In the beginning God created the heavens and the earth.	10	23.8
RU	В начале сотворил Бог небо и землю.	7	16.7
ES	EN el principio crió Dios los cielos y la tierra.	10	23.8
AR	في البدء خلق الله السموات والارض.	6	14.3
FR	Au commencement Dieu créa les cieux et la terre.	9	21.4
TOTAL		**42**	**100.0**

Parallel documents (verses)	31,226
Languages with translations	2,426
Electronically-available parallel translations	80+
Coverage of internet content	99.75%

Sandia National Laboratories

Potential 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**

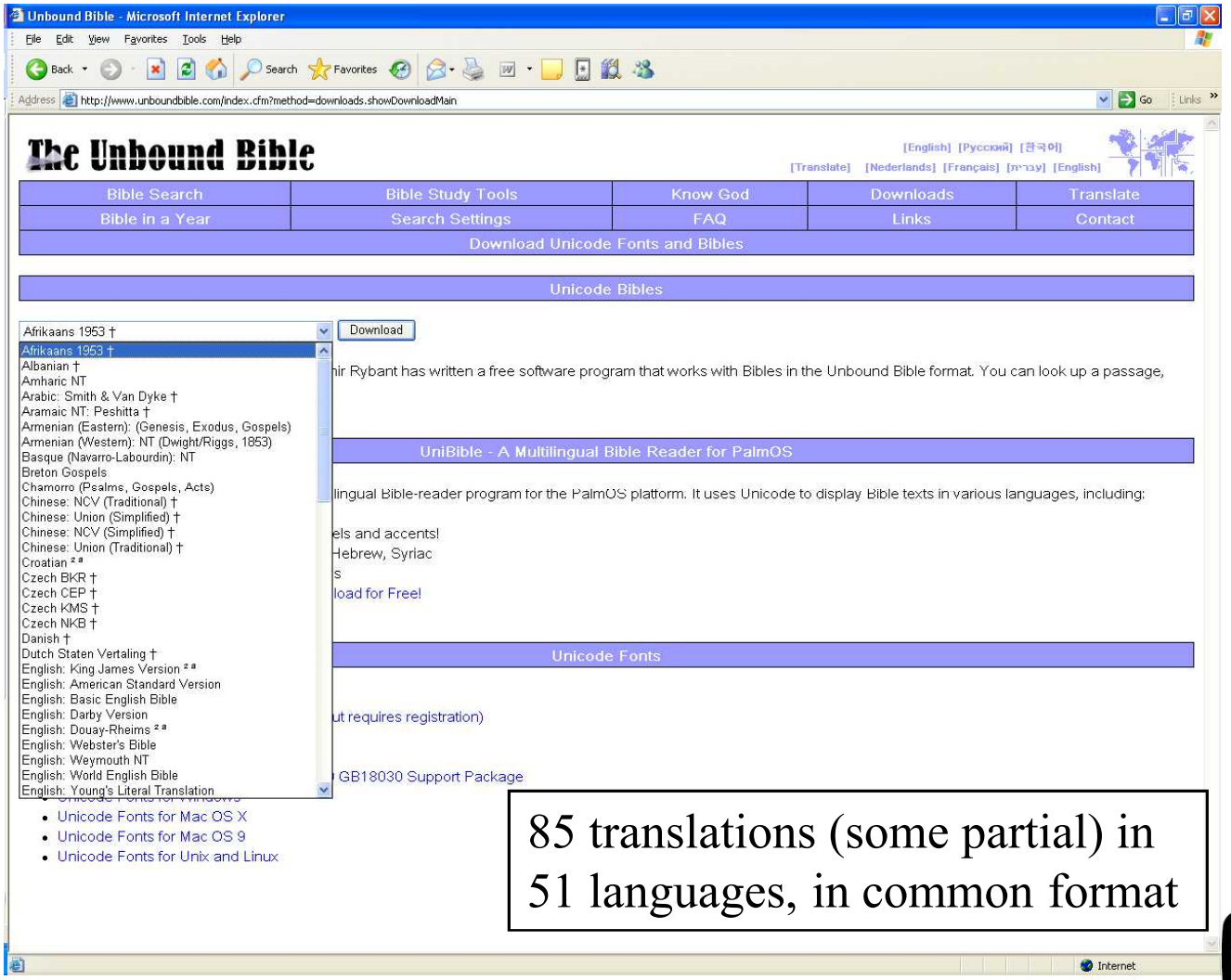
* This column is a sub-section of the Bibles column – for example, there is a translation of the Deuterocanon for 47 of the 61 languages of Europe in which the Bible has been translated.

[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

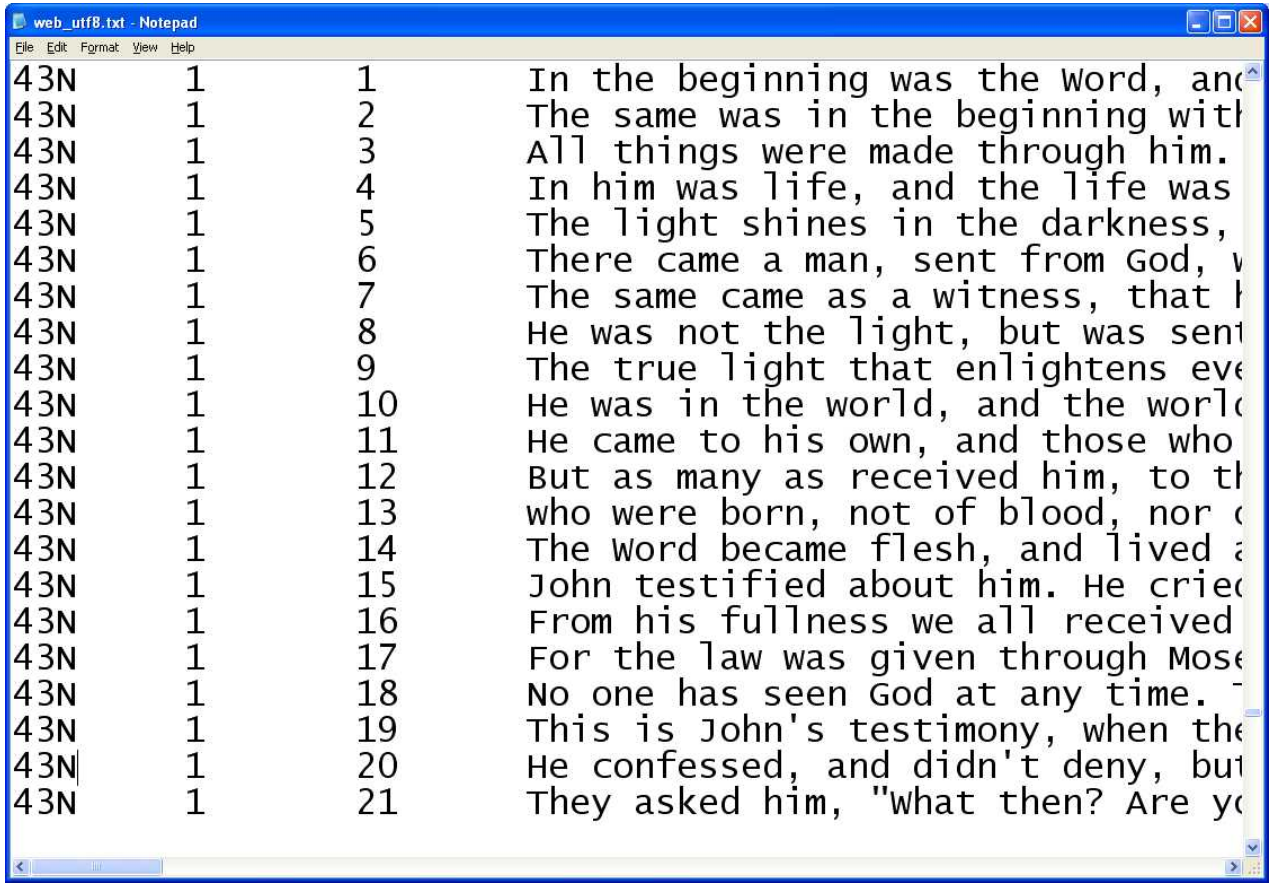
The ‘Unbound Bible’



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

Sandia National Laboratories

The ‘Unbound Bible’ – a sample

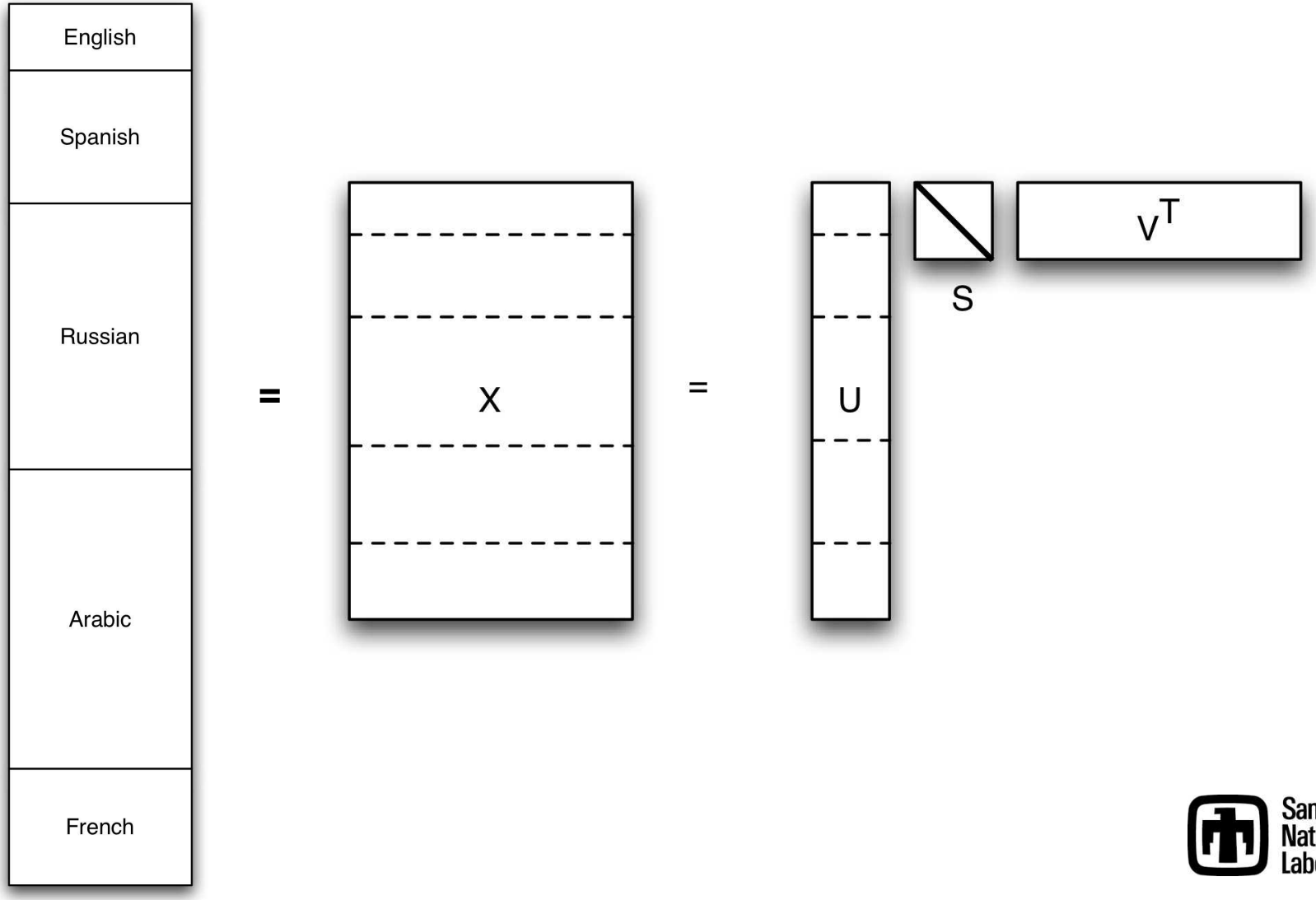


Sandia National Laboratories

Languages in our implementation

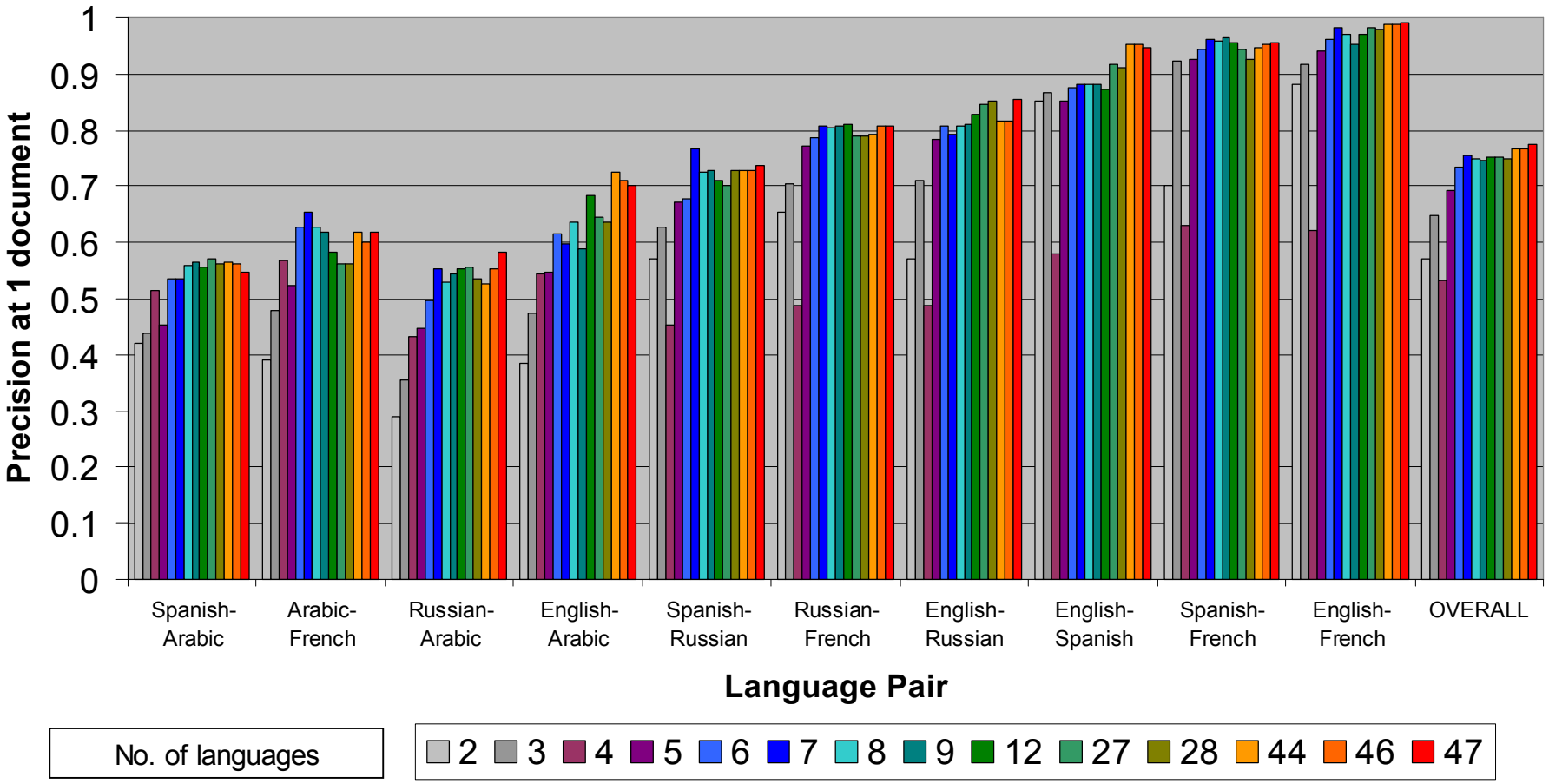
Afrikaans	Estonian	Norwegian
Albanian	Finnish	Polish
Amharic	French	Portuguese
Arabic	German	Romani
Aramaic	Greek (New Testament)	Romanian
Armenian Eastern	Greek (Modern)	Russian
Armenian Western	Hebrew (Old Testament)	Scots Gaelic
Basque	Hebrew (Modern)	Spanish
Breton	Hungarian	Swahili
Chamorro	Indonesian	Swedish
Chinese (Simplified)	Italian	Tagalog
Chinese (Traditional)	Japanese	Thai
Croatian	Korean	Turkish
Czech	Latin	Ukrainian
Danish	Latvian	Vietnamese
Dutch	Lithuanian	Wolof
English	Manx Gaelic	Xhosa
Esperanto	Maori	

The standard multilingual LSA model



Sandia National Laboratories

‘Massive linguistic parallelism’ helps



Sandia National Laboratories

Technical challenges in multilingual clustering

- With MLSA, documents do not cluster language-independently

Ranking	Language of Retrieved Document	Relevant?
1	English	✓
2	English	✘
3	English	✘
4	English	✘
5	English	✘
6	French	✓
7	Spanish	✓
8	Arabic	✓
9	Russian	✓

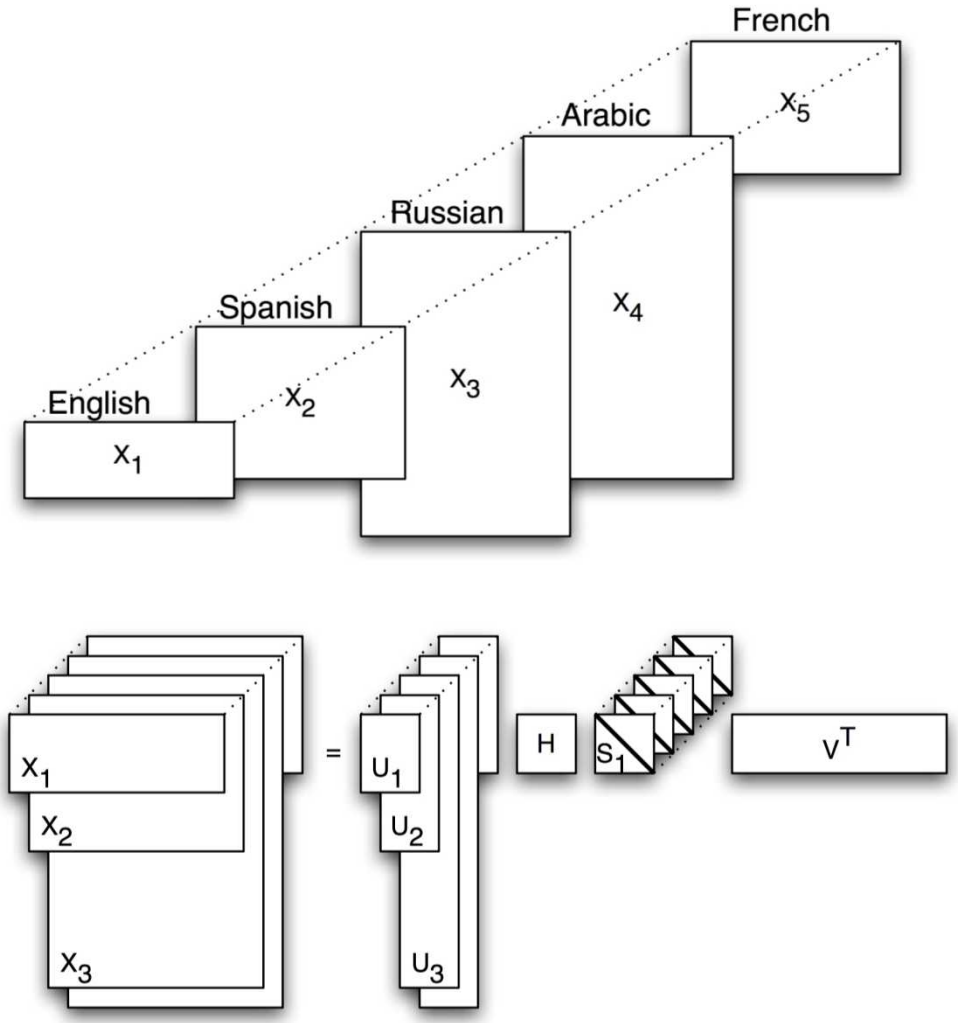
Why standard LSA has drawbacks

Different languages have different statistics

Language	Types	Tokens
English	12,335	789,744
Russian	47,226	560,524
Spanish	28,456	704,004
Arabic	55,300	440,435
French	20,428	812,947

Analytic vs. synthetic languages

The PARAFAC2 model



Sandia National Laboratories

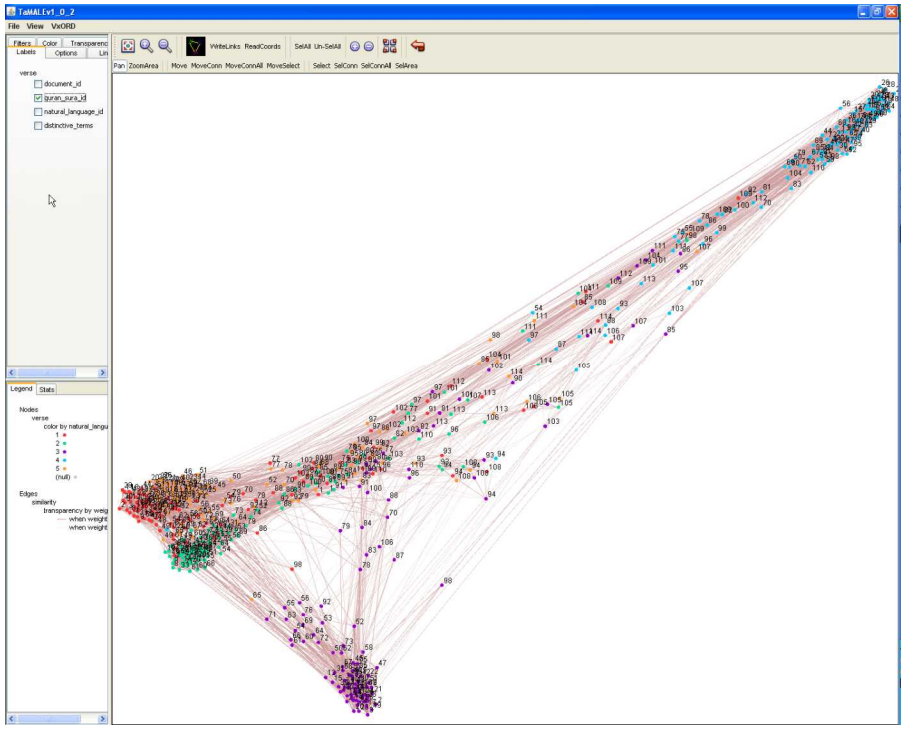
Comparison of LSA and PARAFAC2 (1)

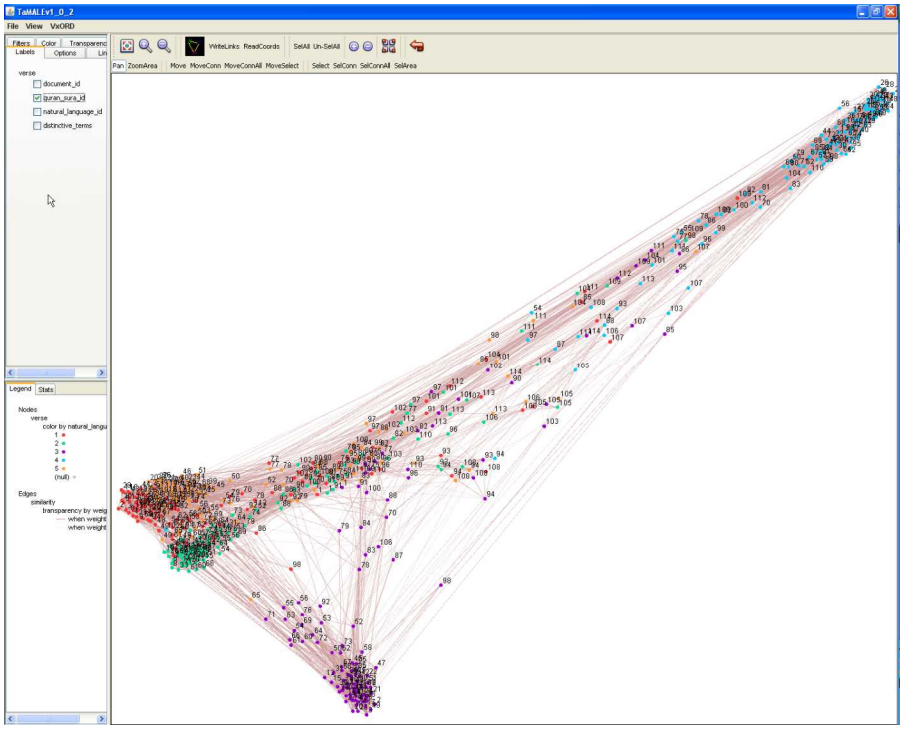
Parameters for common-basis comparison

- 5 parallel languages in training data
- 70 dimensions
- Log-entropy weighting scheme

Common-basis comparison	LSA	PARAFAC2
Multilingual precision at 0	0.3473	0.5722
Multilingual precision at 5 docs	0.3330	0.5568
Precision at 0	0.7190	0.8079
Precision at 1 doc	0.6418	0.7509
Best achieved (>= 5 training languages)		
Multilingual precision at 0	0.3473	0.6504
Precision at 0	0.8819	0.9342

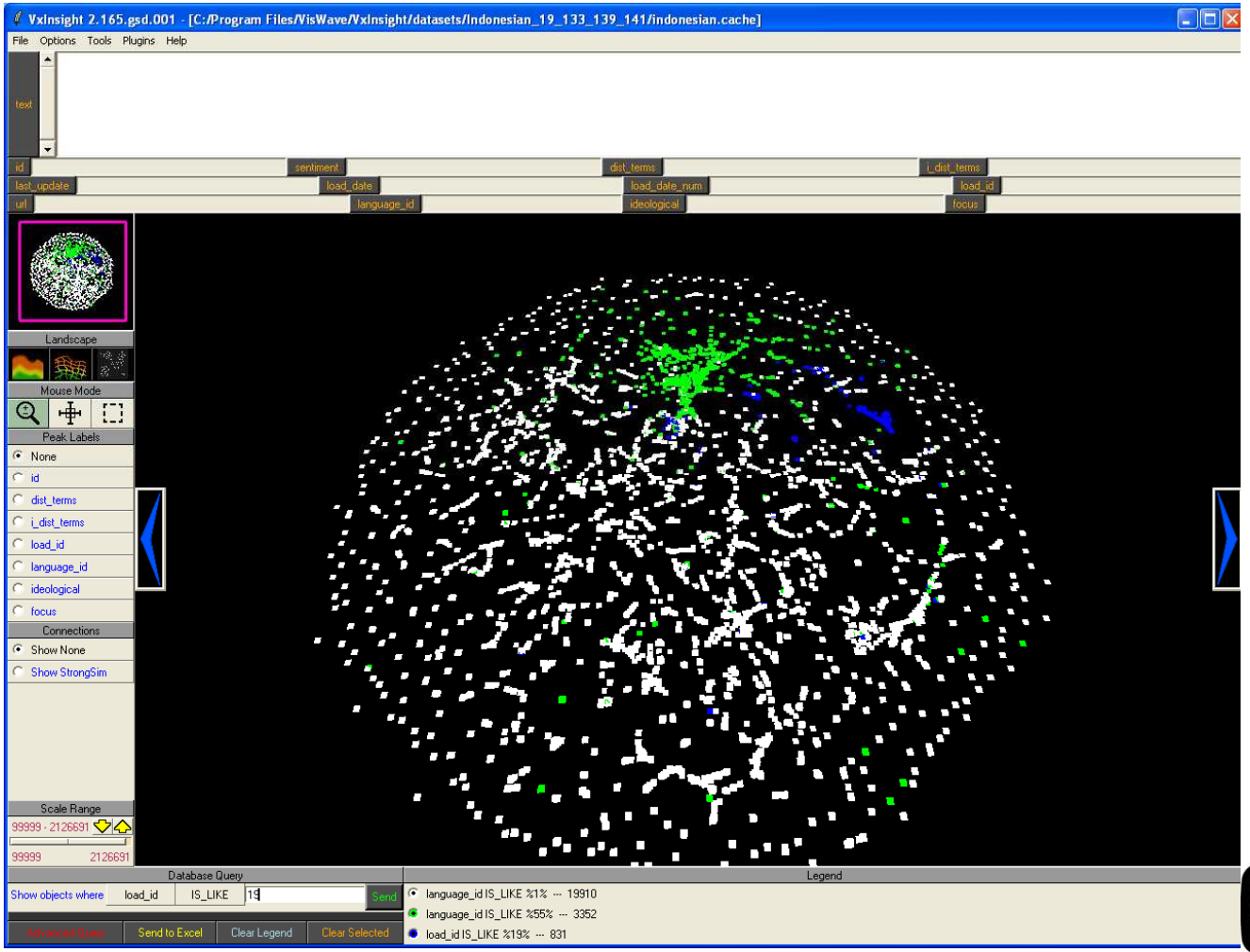
Comparison of LSA and PARAFAC2 (2)





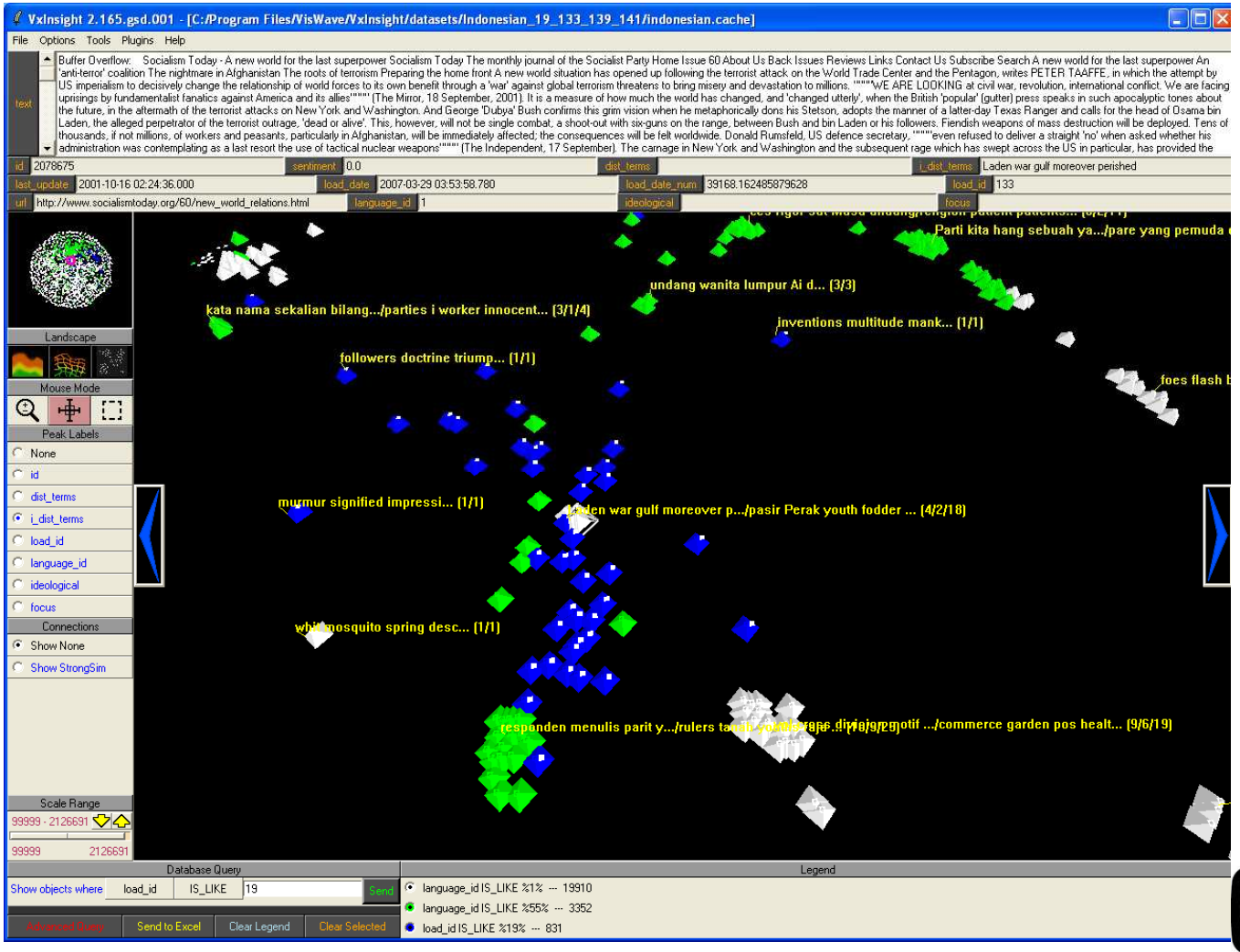
Sandia National Laboratories

Multilingual visualization



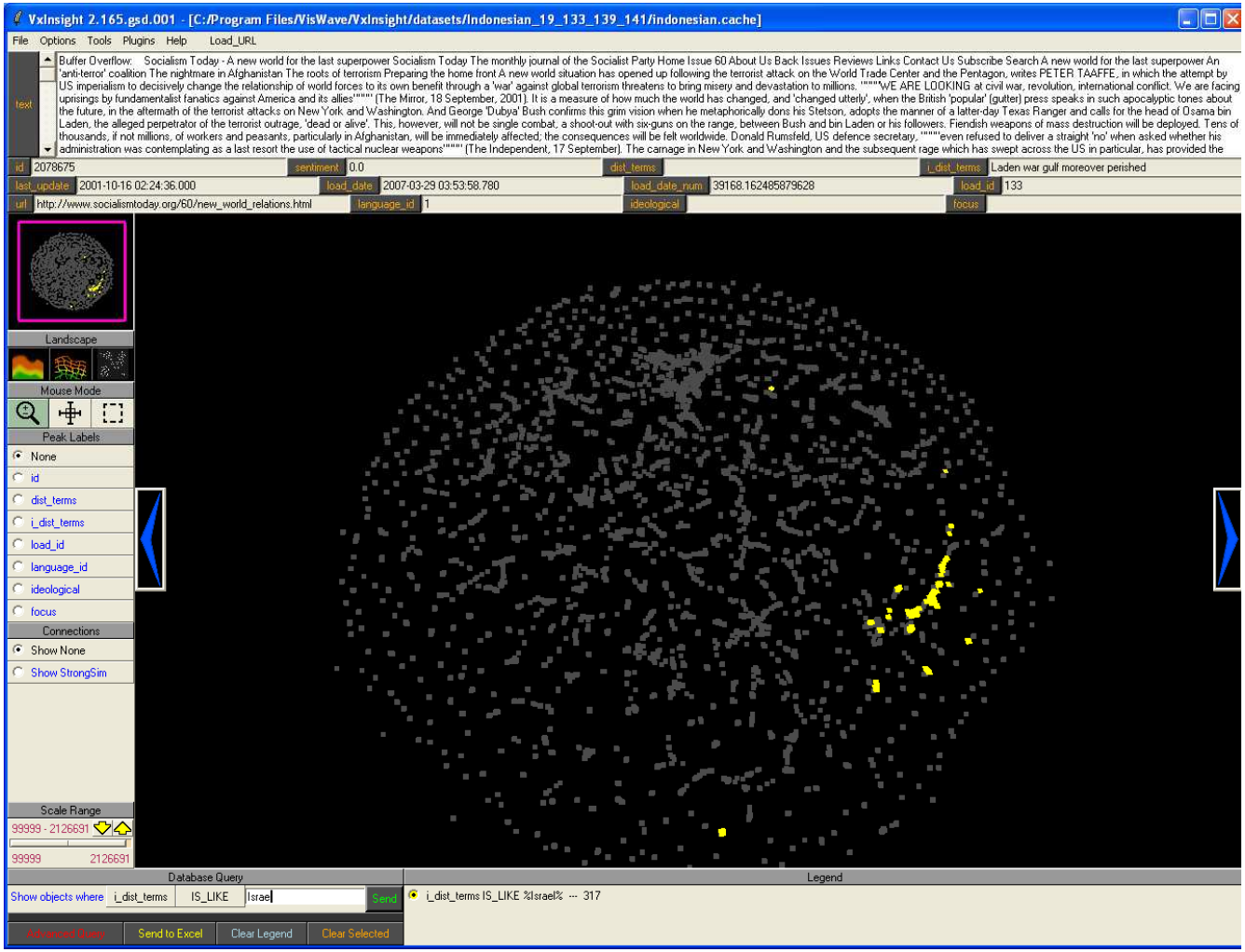
Sandia National Laboratories

Cluster detail



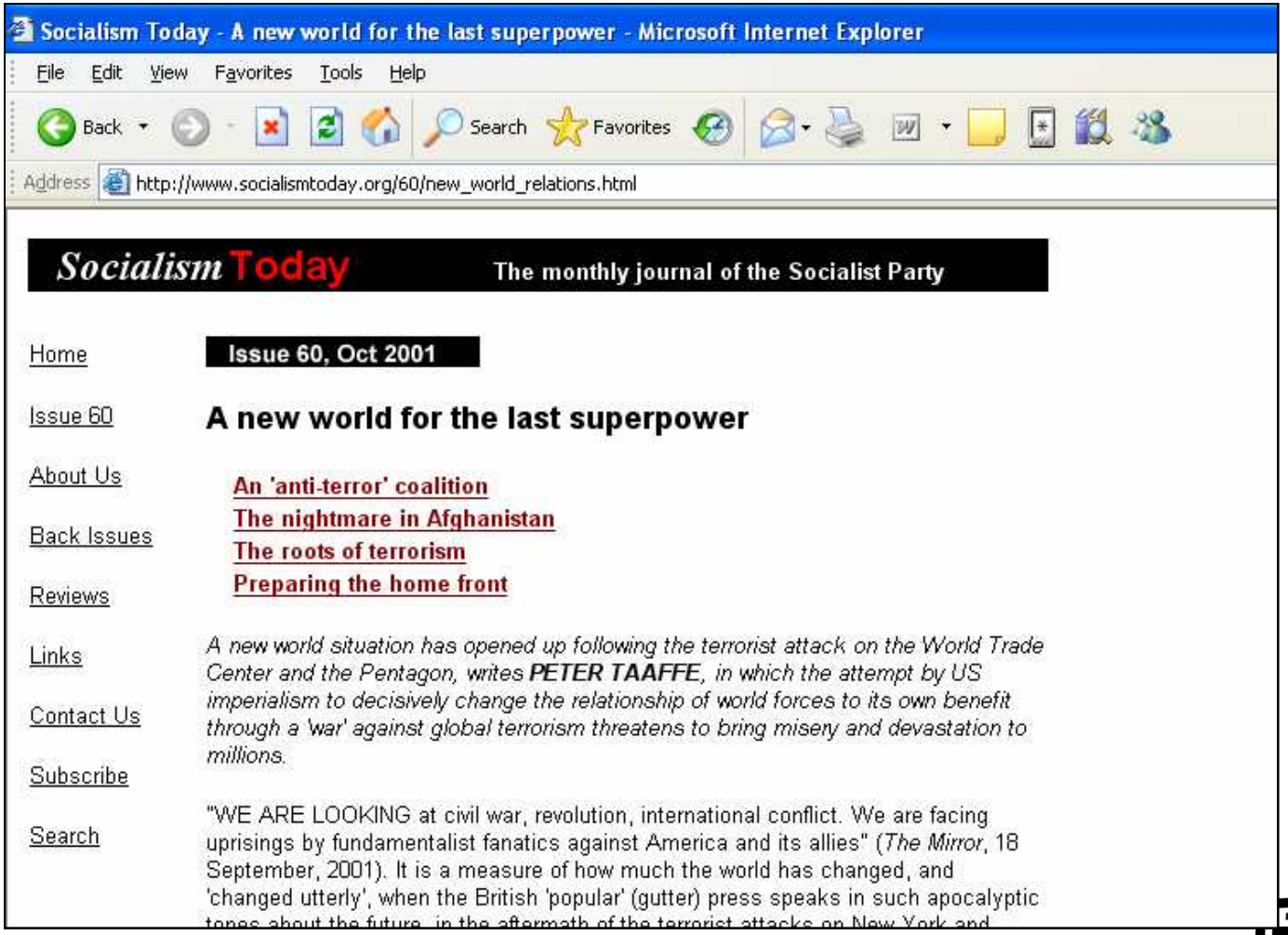
Sandia National Laboratories

Where do topics of interest fit into the overall space?



Sandia National Laboratories

Navigation to documents of interest



Sandia National Laboratories

References and contact information

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

- Chew and Abdelali. 2007, forthcoming. Benefits of the 'massively parallel Rosetta Stone': cross-language information retrieval with over 30 languages, *Proceedings of the Association for Computational Linguistics conference*, 2007.

- Chew, Bader, Kolda and Abdelali. Forthcoming. Cross-language information retrieval using PARAFAC2. Submitted to KDD 2007.

- Chew and Abdelali. Forthcoming. The Effects of Language Relatedness and Morphology on Multilingual Information Retrieval: A Case Study With Semitic Languages

SANDIA POINT OF CONTACT:
Peter Chew (pchew@sandia.gov)

Sandia National Laboratories

DISCUSSION
and
QUESTIONS

SANDIA POINT OF CONTACT:
Peter Chew (pchew@sandia.gov)

Sandia National Laboratories