CROSS-LANGUAGE INFORMATION RETRIEVAL:

A RESEARCH ROADMAP

Fredric Gey, Noriko Kando, Carol Peters

Why a Research Roadmap Workshop?

- **CLIR has been around for a decade**
- **Much progress has been**
- **Time to summarize and design our research future**
- **Goal: a five year research and development plan**
- **Influence on program managers and funders**
- **Special issue of IP&M**

CLIR: A RESEARCH ROADMAP

Content

- **Fifteen papers (13 presenters)**
- **Four main papers selected**

1. **(Oard) When you come to a fork in the road, take it: Multiple Futures for CLIR research**
2. **(Nie) Towards a Unified Approach to CLIR and Multilingual IR**
3. **(Jones) CLIR: Consolidating and Moving Forward**
4. **(Mayfield & McNamee) Three Principles to Guide CLIR Research**

CLIR: A RESEARCH ROADMAP

Challenges

- **Where to get resources for resource-poor languages**
- **Why aren't search engines using our research?**
- **Building a web corpus in multiple languages**
  - **Issues: English domination, leading to imbalance**
  - **Character and font representation**
- **Your challenge here …**

CLIR: A RESEARCH ROADMAP
ORGANIZATION

- **Six thematic sessions,**

1. **Approaches to CLIR**
2. **Languages with little resources**
3. **Multimedia**
4. **User Studies/interactive**
5. **Evaluation**
6. **Building a Roadmap**

CLIR: A RESEARCH ROADMAP

1. Approaches to CLIR

- Extensible systems based upon dictionary-based query translation (suitable for languages without MT development)
- Multilingual thesauri (Russian-English) – automated IR use of thesauri requires information not normally present in thesauri designed for manual indexing (argued for corpus-based thesaurus development)
- Pivot languages where translation pairs not available (English-German-Swedish-Finish) extends the range of CLIR language pairs

Oard: In 1996 We Had A New Challenge

- **Community formed around an agreed problem**
  - **Extend ranked retrieval to cross-language search**
- **Initial exploratory work looked promising**
  - **CL-LSI, Radwan&Fluhr, SIGIR 96 Workshop**
- **Most urgent need was for test collections**
  - **TREC (97), TDT (98), NTCIR (99), CLEF (00)**

Oard in 2002: CLIR is a Solved Problem!

- **Nearly 100% of monolingual effectiveness**
  - **Robust translation, natural expansion effect**
- **Small bag of tricks work across many languages**
  - **Stemming, term selection, weight mapping**
- **Adequate resources for many language pairs**
  - **Term lists, monolingual corpora, parallel Web text**

OARD: So Why Aren't People Using It?

- **#3: Genre**
  - **We know lots about news, little about Web pages, scientific-medical document collections**
- **#2: Efficiency**
  - **Little work on indexing-time approaches**
- **#1: Utility**
  - **How will the ranked list be used?**

CLIR: A RESEARCH ROADMAP

WHAT HAVE WE LEARNED IN SIX YEARS?

- **Good monolingual retrieval essential**
  - **Stemmers, stopword lists**
  - **Decompounding**
  - **Morphology?**
  - **Segmentation for character languages**
  - **N-grams can rival by-word retrieval**
- **Phrase translation important**
- **Blind feedback works well across languages**
- **All this has been tested by evaluations**

CLIR: A RESEARCH ROADMAP

Gareth Jones: Where are we now?

- **Bilingual, Multilingual CLIR works**
- **Non-English Monolingual IR - local expertise important**
- **Would be good for progress of CLIR overall if we could share**
- **Need more resources and expertise.**

- **Publication of work in CLIR is very distributed - hard to develop a solid understanding of what really works best.**
- **Need comparative analysis highlighting areas needing more**
- **further experiments needed, more training data, etc.**

J-Y Nie: **Current approaches to multilingual IR**

1. Query translation
2. Monolingual retrieval
3. Result merging

(Even for a mixed collection)

- Steps 1 and 2 for each language separately
- Step 3 considers the languages together

- Relationships between languages: weighting, retrieval
- Should the relationships between languages considered earlier?
- Should the collections merged earlier?

Nie: Problems in the current approaches

1. **Translation is governed by general translation tools**

- **Likely to select the most common translation words**
  - **MT**
  - **Dictionary**
  - **Parallel corpus**
  - **Reasonable for general purposes**
  - **Is it also reasonable for IR?**
  - **Not necessarily: discriminative words**
    - **E.g. drug vs. narcotics for "drug traffic"**
  - **Corpus information helps, but not definitive**
- **Domain independent translation**
  - **Making translation dependent on the collection**

Nie's proposals

- **A unified model integrating translation and retrieval, which takes into account:**
  - **Term frequency**
  - **Document frequency**
  - **Language share in the mixed collection**

Treat CLIR as a special case of inferential IR

- **IR is an inference process: Can we infer a query from a document?**
- **Translation is a step in the inference**
- **Similar to query expansion (infer related terms)**

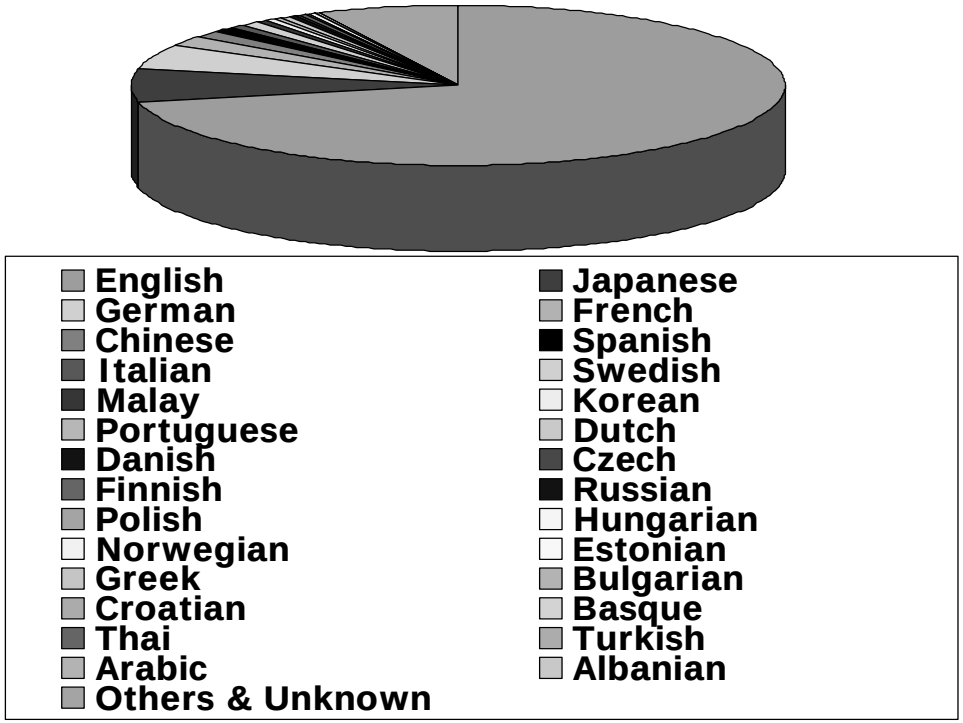
Nie's Proposals (Continued)

2. **Experiments on realistic mixed collections (e.g. a Web collection ?)**
   - **Investigate the retrieval in a mixed collection**
   - **Develop appropriate weighting schemas in indexing to better coordinate languages (instead of merging afterwards)**

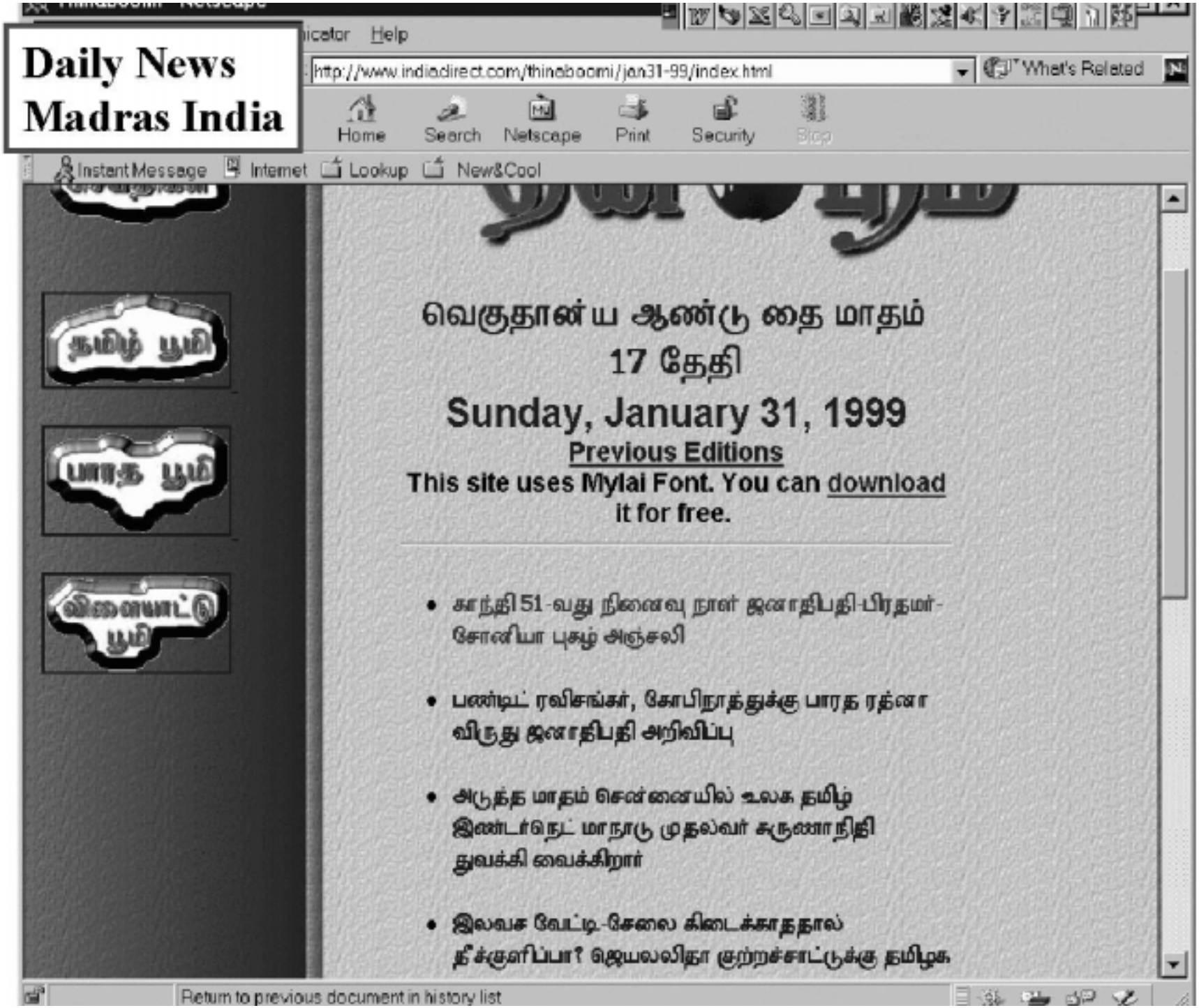
1999 Language Distribution on a 634 Million Web Pages Corpus

Language	Number of Docs	Percentage
English	453,685,690	71.5288%
Japanese	43,271,080	6.8222%
German	32,253,563	5.0851%
French	11,107,994	1.7513%
Chinese	9,642,450	1.5202%
Spanish	6,965,560	1.0982%
Italian	5,638,827	0.8890%
Swedish	4,392,709	0.6926%
Malay	3,619,227	0.5706%
Korean	3,200,762	0.5046%
Portuguese	3,014,294	0.4752%
Dutch	2,745,610	0.4329%
Danish	1,911,677	0.3014%
Czech	1,428,385	0.2252%
Finnish	1,312,932	0.2070%
Russian	1,150,127	0.1813%
Polish	952,716	0.1502%
Hungarian	760,162	0.1198%
Norwegian	607,211	0.0957%
Estonian	456,613	0.0720%
Greek	393,360	0.0620%
Bulgarian	392,777	0.0619%
Croatian	310,237	0.0489%
Basque	258,074	0.0407%
Thai	99,691	0.0157%
Turkish	81,218	0.0128%
Arabic	38,167	0.0060%
Albanian	17,779	0.0028%
Others & Unknown	44,561,062	7.0256%
Total	**634,269,953**	**100%**

Language Distribution of Web Content



Daily News
Madras India



CLIR: A RESEARCH ROADMAP

3. CLIR for Multimedia

- **Sanderson EuroVision**
  - **Recommends that Image (and other multimedia) retrieval) be considered for CLIR, because the retrieved objects can be understood independent of language**

CLIR: A RESEARCH ROADMAP
3. CLIR for Multimedia (Gareth Jones)

- **Many new tasks being investigated in IR: multimedia,**
- **question-answering, summarisation, web-retrieval.**
- **– All of these could be extended to cross-lingual or multi-lingual**
- **tasks.**

- **Are cross-lingual or multi-lingual versions of these tasks**
- **worthwhile?**
- **– If so, which order should they be taken in?**

- **Are the monolingual (English!) versions of these technologies**
- **sufficiently mature to make cross-lingual or multi-lingual versions**
- **worthwhile at the present time?**

CLIR: A RESEARCH ROADMAP

3. CLIR for Multimedia (Gareth Jones)

- **Indexing multimedia content requires often scarce or expensive technology.**
- **e.g. a high quality speech recognition system for each language and enough spare computing power to transcribe a potentially very large amount of data.**

- **TREC Spoken Document Retrieval and Video tracks provide**
- **baseline indexing data or share data among participants.**
- **– but this assumes that someone associated with the task has the requisite technology!**

CLIR: A RESEARCH ROADMAP

4. User Studies and Interactive CLIR

- **Where are the postulated monolingual searchers – in Europe many users are polyglot (Petrelli)**
- **English is used as slang in other languages – should be incorporated into extended search engines (Petrelli)**
- **Users adapt to systems capabilities easily, but not to the document language(Gonzalo)**
- **Proposal: Interactive Cross-Language question answering as more realistic than monolingual question answering (Gonzalo)**

Petrelli: A Surprise Result?

- **Polyglots are the majority of potential (European) users**
- **Users search for use: search skills very low**
- **Search many languages simultaneously**
- **Swap between the known languages**
- **English used as pivot (multi-language query)**
- **Needs to search phrases and proper names**
- **User-created dictionaries**

Petrelli: The 3 key points for CLIR user-centred research

- **Understand**
  - **Users, uses, and environments**
- **Design**
  - **System functionalities**
  - **Interface features**
- **Evaluate**
  - **Different interface layout**
  - **Different CLIR techniques**
  - **Cognitive impact and workload**

Evaluation -- the Way Ahead
A Case of the NTCIR

Noriko Kando

National Institute of Informatics (NII), Tokyo

kando@nii.ac.jp

Implications of Evaluation Workshops

- large-scale test collections,
- research idea exchange and technology transfer
- "showcase" of the new technologies
- motivation of research
- discussion on evaluation methods
- the model of experimental design
- attracting newcomers and so on.

Axes to characterize CLIR systems

- Languages
- Type of media
- Tasks and users
- Relevance judgments or success criteria
- Document genres
- Layers of CLIR technologies
- Information access process

Future Directions

- Cumulating the experiences for each language and each language pair
- Switching (pivot) language CLIR
- Task/genre oriented CLIR
- Pragmatic layer of CLIR technologies and identifying the differences
- Towards CL information access – QA, Summarization, text mining, identifying the differences of the viewpoints across the languages or cultures

Carol Peters: CLEF Evaluation Questions:

- **What is the relationship between evaluation campaigns and CLIR system development?**
- **Are evaluation campaigns doing enough?**
- **If not, what should they be doing?**
- **How can evaluation campaigns be supported?**
- **Does current research in CLIR reflect the needs of the application communities?**

CLEF (and TREC) Six Years of Activity

Focus on text retrieval

- monolingual/bilingual/multilingual free text retrieval tasks
- mono- and cross-language IR on structured data

Growth in participation

- 13 groups in 1997 – ca 40 groups in 2002
  - more European groups – more industrial groups
- annual workshops

Creation of test collection

- comparable corpus in 8 languages; queries in 12
- scientific texts collection in German and French
- data and relevance assessments from past campaigns are available to registered participants free-of-charge

CLEF: Points for Discussion

- how do we meet the needs of all (European) languages
- what type of coordination and funding model should we be adopting ?
- how far can we go with limited funds and much voluntary work?
- how can we make our test-suites publicly available at a reasonable cost?
- what new tasks/evaluation methodologies are needed to address more advanced information requirements?
- how can we best reduce the gap between research and application communities?

CLIR: A RESEARCH ROADMAP

2. Languages with Few Resources

Pirkola: Zulu as a case language with few resources

- **Increasing amounts of information is available in indigenous languages**
- **Indigenous languages exacerbate the challenge of conceptual mismatch which limits the utility of lexical mappings**

CLIR: A RESEARCH ROADMAP

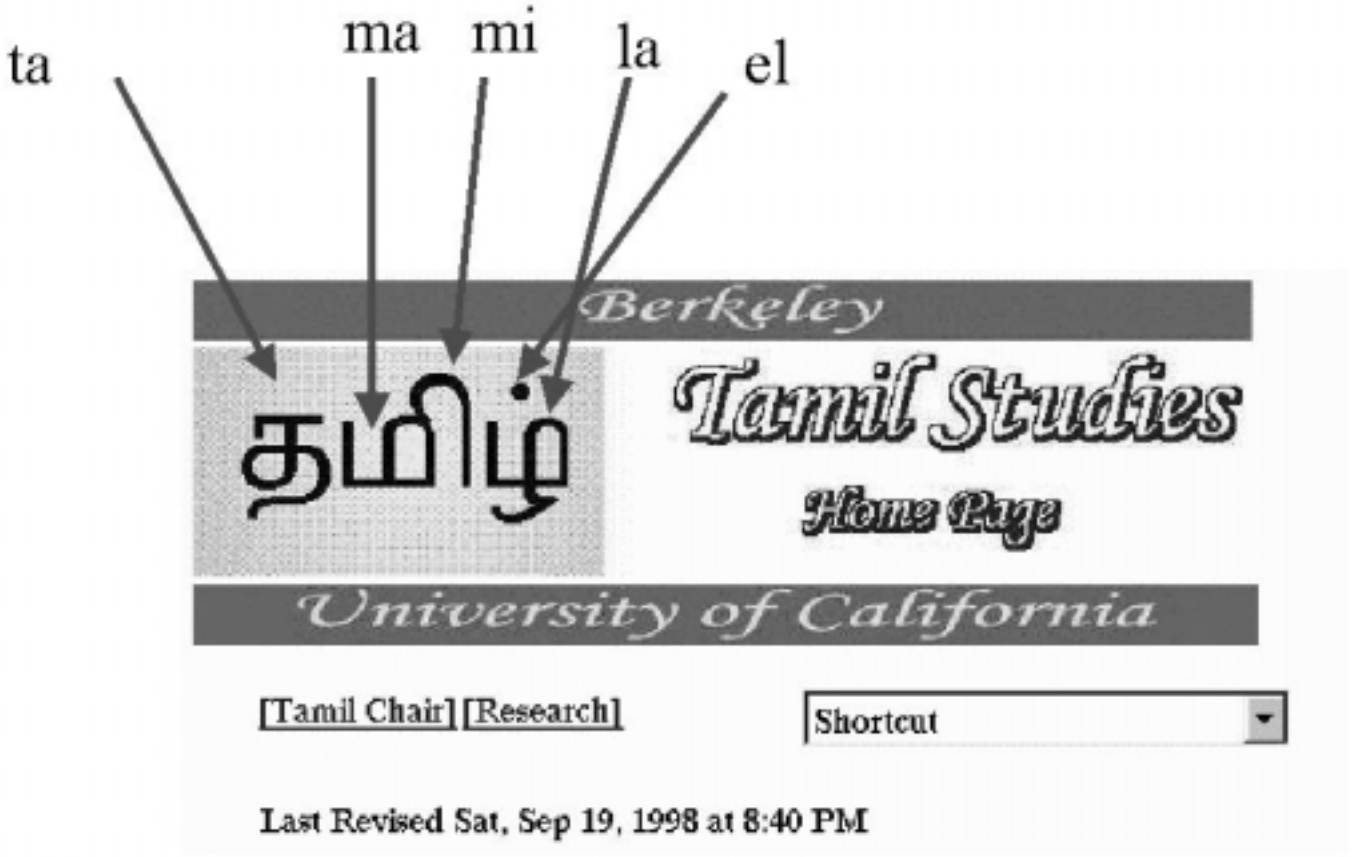
2. Languages with Few Resources

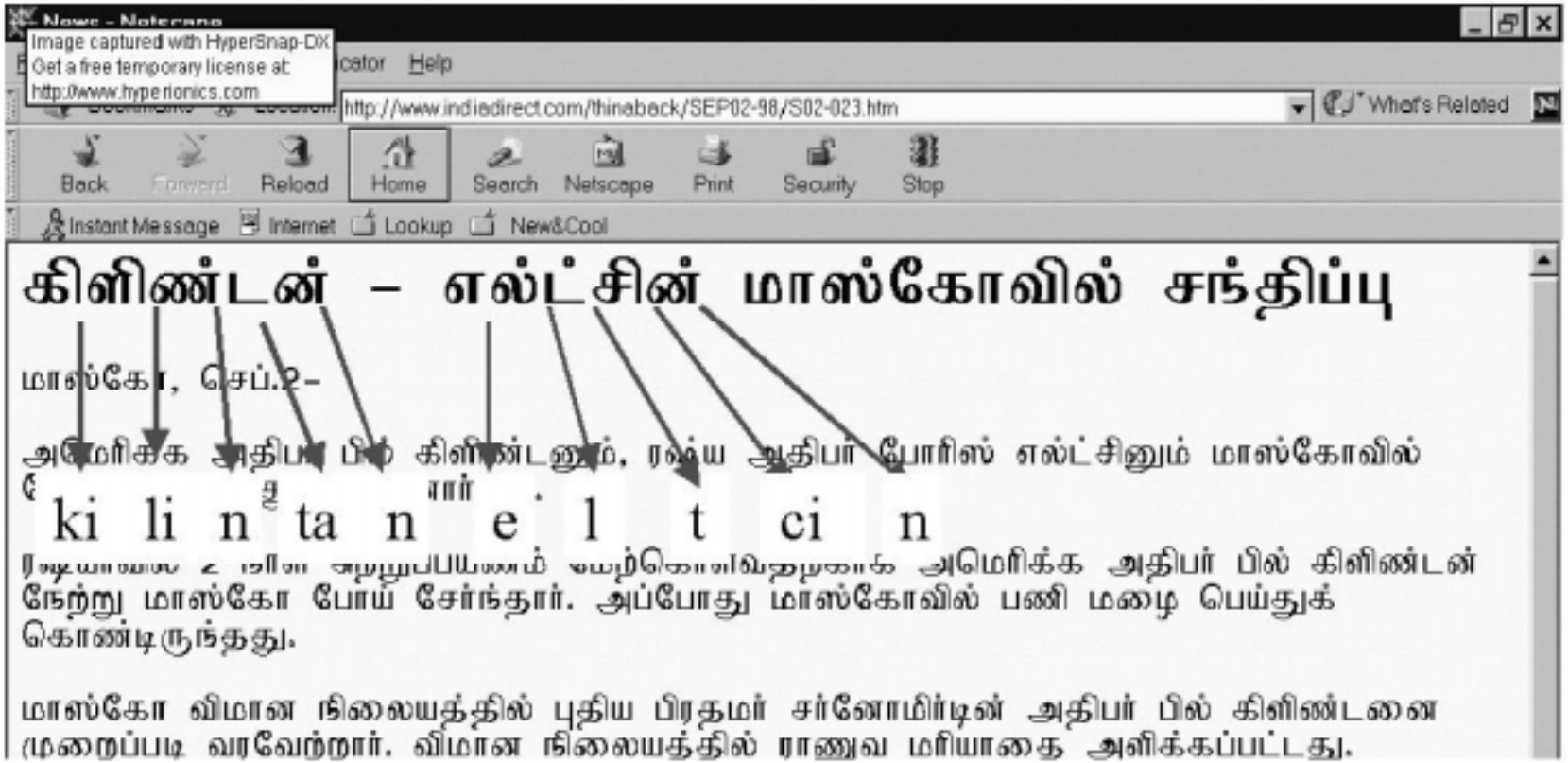
Gey: STARTING FROM NOTHING: Resources of First Resort in CLIR

- **Six of the top 25 languages by population speaking are from the Indian Subcontinent (Hindi, Bengali, Urdu, Telugu, Punjabi, Tamil). Virtually nothing has been done (in the west) with these languages.**
- **Gey argues that even in the absence of resources something can be done: transliteration and phonetic recognition can combine to supply primitive retrieval**

How can we deal with these languages with no resources?

- **We look for a search mechanism which needs no (or little) resources**
  - **submit our query in English**
  - **Search the target language as miss-spelled English**
  - **Buckley in TREC-6: English-French**
- **Amaryllis task in CLEF-2002**
  - **Amaryllis thesaurus – 173,946 English-French term pairs**
  - **Accumulation capital – Capital Accumulation**
  - **Exact matches (case normalized): 44,975**
  - **Exact matches after French accents normalized: 49,926**
  - **Above techniques with word permutation: 52,790**
  - **Biological accumulation – Accumulation biologique**
- **General idea: A lot of borrowed words are already out there in the target language**


ta
ma
mi
la
el
Berkeley
தமிழ்
Tamil Studies
Home Page
University of California
[Tamil Chair] [Research]
Shortcut
Last Revised Sat, Sep 19, 1998 at 8:40 PM



Clinton – Yeltsin meeting in Moscow

Moscow, Sep. 2-
America's president Bill Clinton and Russia's president Boris Yeltsin met yesterday in Moscow and talked. America's president Bill Clinton arrived yesterday in Moscow to undertake a 2-day tour in Russia. Then it was snowing in Moscow. At Moscow's new airport the premier minister Sarnomirtin welcomed President Bill Clinton in the proper way. The army paid respect at the airport. At that time both national anthems were played.
(translation by Arash Zeini, University of Cologne)

ரஷ்ய அதிபர் போரிஸ் எல்ட்சின் பிரதமர் கிரியங்கோவை பதவியிலிருந்து நீக்கப்பட்டு புதிய பிரதமராக சர்னோமிர்டினை நியமித்துள்ளார். ஆனால் இந்த நியமனத்தை கம்யூனிஸ்ட்டுகள் அதிகம் இருக்கும் பாராளுமன்றம் நிராகரித்துவிட்டது.

The no-resource solution: transliteration + phonetic matching

- **Gadd's PHONIX algorithm**
- **Pre-process text to substitute characters**
- **Remove vowels**
- **Match on consonants (used for spelling correction)**
- **Economic policy → ekonomik polici → eknmk plc**
- **Ekonomicheskiaa politika → eknmck pltk**
- **Edit distance match: 3**

CLIR: A RESEARCH ROADMAP
6. A Research Roadmap

- **Three Principles to guide CLIR Research (Mayfield and McNamee)**
  - **CLIR = CL X IR (translation is central to CLIR)**
  - **CLIR > CLDR**
  - **MLIR != BLIR**
- **Five Dangers**
  - **Perceived barriers to entry**
  - **Availability of language resources**
  - **Waning interest among funders**
  - **Unclear path to Usefulness**

CLIR: A RESEARCH ROADMAP

6. A Five Year Roadmap for CLIR Research (Mayfield and McNamee)

1. **Resources: Standards and tools for translation resources Ml/BL.**
   Evaluation: Isolation of resources from retrieval methodology
2. **Resources: Large comparable/aligned corpora for several genres.**
   Evaluation: Eval name xlation/tranlit, spell corr for 5-10 languages
3. **Resources; Construction of BL dicts in >3 languages >100K entries.**
   Evaluation: Document selection by users w/o ability to read language of document
4. **Resources: Tools for building speech models from spoken corpora.**
   Evaluation: Multilingual retrieval in >15 languages (Asia, EU,Indic and Semitic languages)
5. **Resources: WordNet in 15+ languages with core 100K synsets/lang.**
   Evaluation: CL Speech Retrieval in >=4 languages with >10 groups

CLIR: A RESEARCH ROADMAP

6. Final Discussion Points

- **Have we solved the CLIR problem?**
- **Have we defined the boundaries of the CLIR problem?**
- **Need strategies to bridge betw R&D and applications**
- **Do we need better understanding of more realistic requirements of real users?**
- **What are the strategies for moving forward?**
- **Need a reliable, scalable demonstration system**
- **CL applications can include Q&A, Text categorization, image retrieval (systems may be less robust than text retrieval)**

CLIR: A RESEARCH ROADMAP

6. Where to go for more information

- **Workshop web site:**
  - **http://ucdata.berkeley.edu/sigir-2002**
- **Contains all the workshop papers**
- **Will soon contain all workshop presentations**
- **Will have follow-on news of CFP and workshop summary**

Proposal for a Special Issue on CLIR
Information Processing & Management

Target:

To be a good reference issue for CLIR

Provisional Schedule:

Mar. 2003:Deadline for paper submission

June 2003:Notification of acceptance

Sept 2003:Publish


Plutonium
In Arabic
Chinese
Greek
Japanese
Korean
Russian
Tamil
بلوتونيوم عنصر فلزي
钚
πλουτώνιο
プルトニウム
플루토늄
ПЛУТОНИЙ
புளுதுதோநஜியாம்