



**URSA**

Unicode Retrieval System Architecture

# **URSA: Using a Unicode Toolkit for Cross-Language Text Retrieval and Visualization**

## **TIPSTER Phase III**

### **Mark Davis & Bill Ogden**

<http://crl.nmsu.edu/Research/Projects/tipster/ursa>



URSA

Unicode Retrieval System Architecture

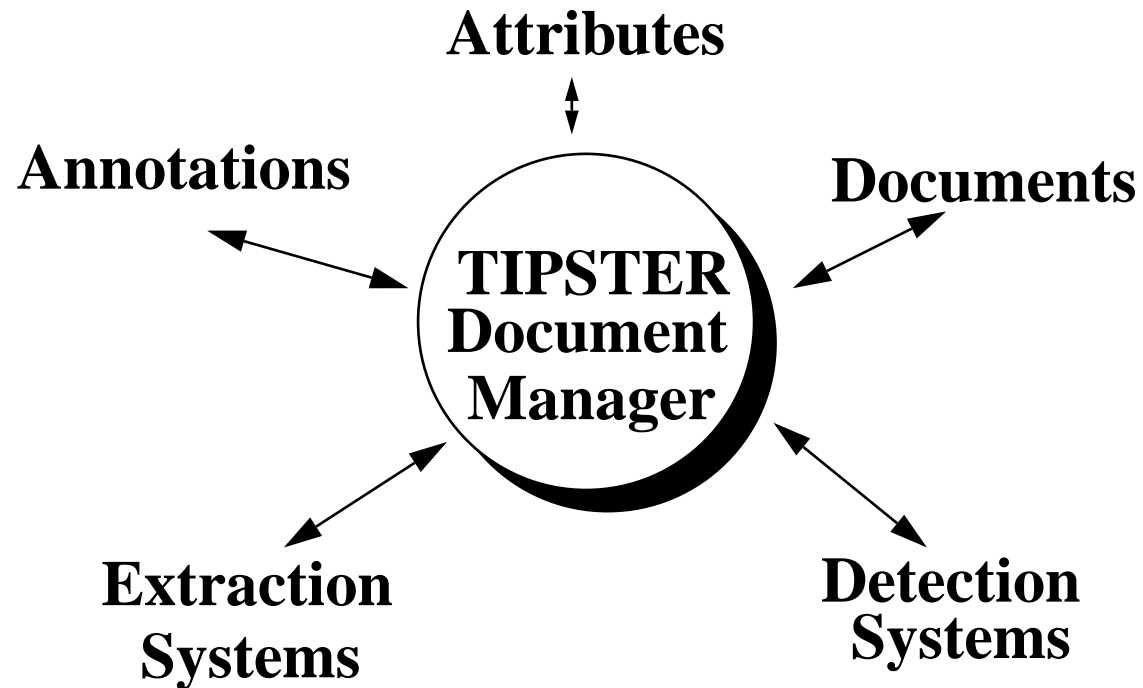
## Multilingual Computing for a Multilingual World

hupou OKTOBER 十月  
NEWS derichos 戦争 冷たい 戦争  
包圍戰



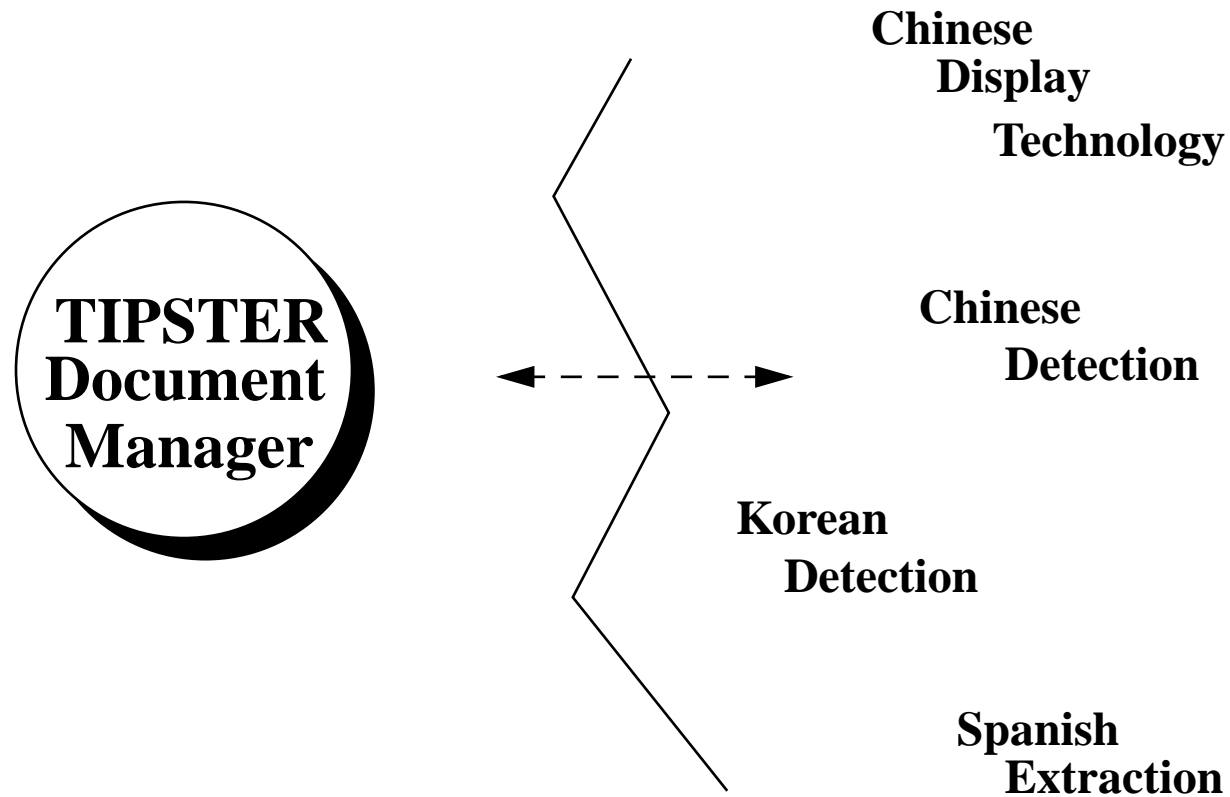


## **TIPSTER can help...**





**...But multilingual computing is still handled outside the architecture by most systems**

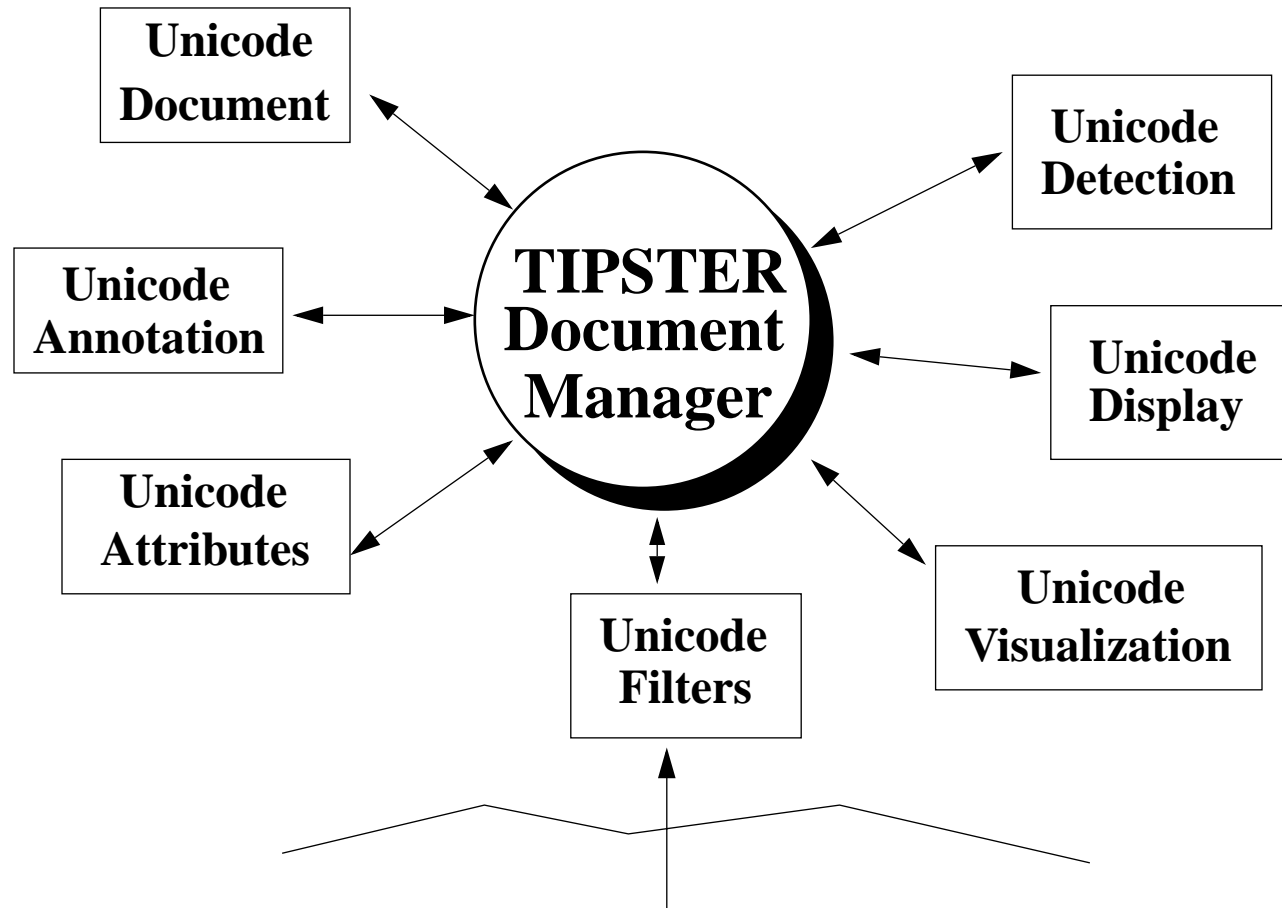




**URSA**

Unicode Retrieval System Architecture

## **URSA (Unicode Retrieval System Architecture) brings true multilingual detection, display and visualization to the TIPSTER architecture**





## Milestones Achieved

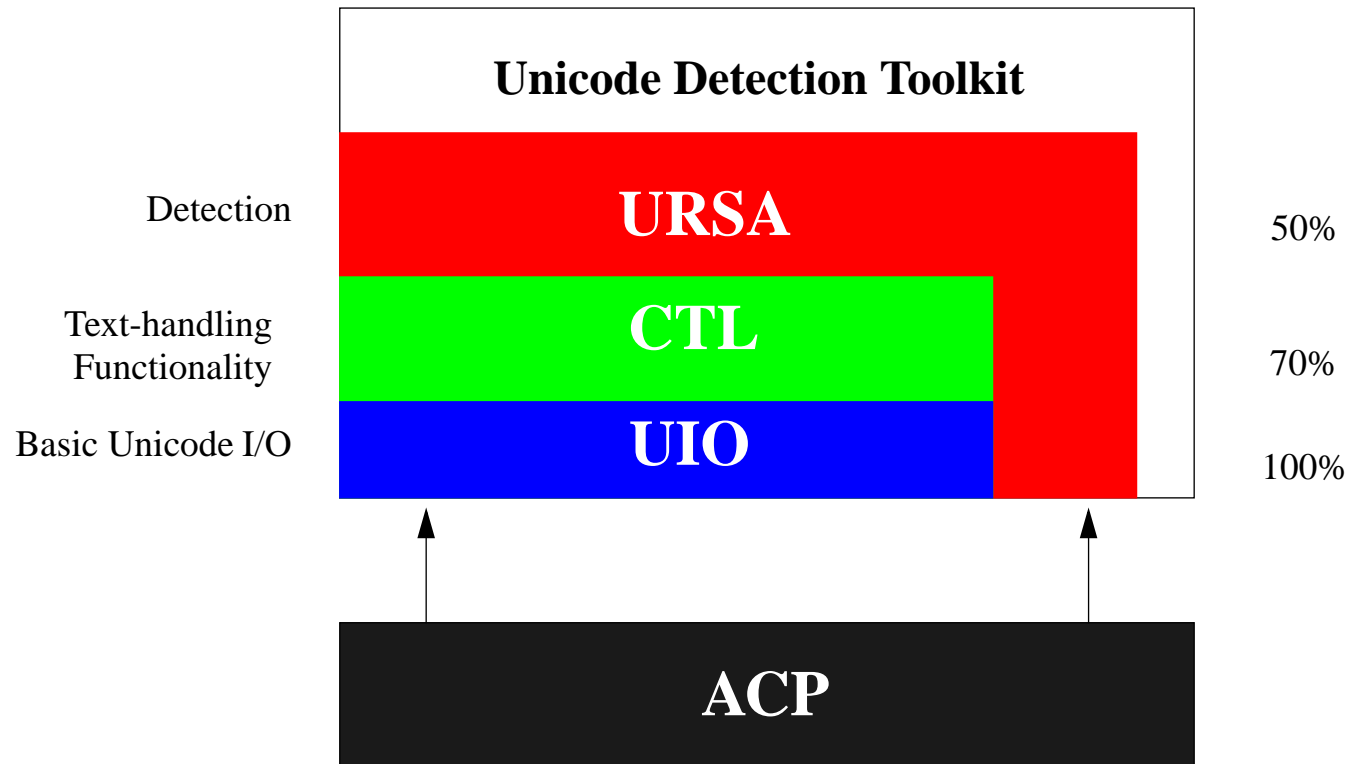
- Under contract as of 8/97!
- Progress on Unicode detection toolkit
- Obtained new, free language resources
- Expanded parallel-text translation methods
- Experimented with Pathfinder networks for visualization and retrieval
- Submitted cross-language TREC runs using English queries against a French database.



**URSA**

Unicode Retrieval System Architecture

## Progress on Unicode detection toolkit





## Common Text Library

- Hash tables
- Lexicons and Postings
- String vectors
- Stemming algorithms (E/F/S)
- Simple SGML parsers
- Phrase extractors (E/F/S)



## Obtained new, free bilingual resources

### Bilingual Lexicons

<i>Languages</i>	<i>Unique Headwords</i>
English-Afrikaans	3,733
English-Dutch	9,853
English-Danish	3,715
English-Finnish	2,832
English-French	3,582
English-Japanese	176,528
English-Hungarian	2,479
English-Italian	2,912
English-Portuguese	2,637



## Developed New Parallel Text Translation Methods

- Subsentence alignment algorithm
- Statistical alignment of French-English parallel texts (Hansards, obtained from Canadian Government WWW site)
- Phrase extraction using cognate anchoring and phrase heuristics.

ENGLISH: Articles discussing the use of sex education to combat AIDS.

DICT: article discutant use sexe éducation combat aids

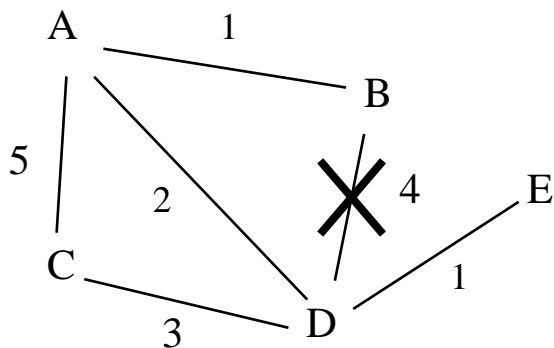
CORPUS: articl discuss caractéristiques enfant scolaire combattre sida

BOTH: article discutant use sexe éducation combat sida

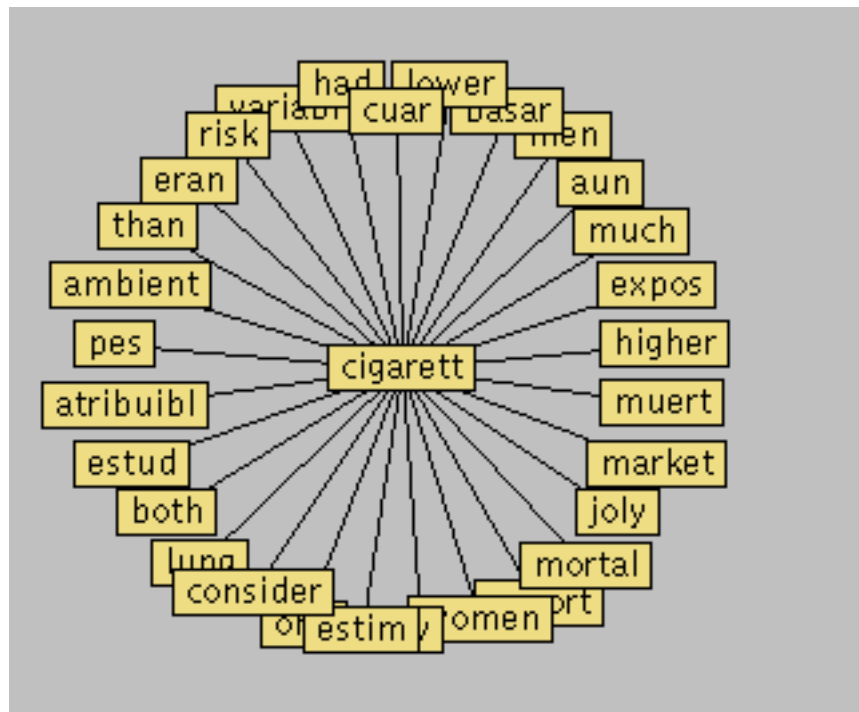


# Experimented with PATHFINDER networks for cross-language visualization

A	B	C	D	E
	1	5	2	
		2	4	
			3	
				1



Shorter 2-link paths beat single links!





## The Next 12 Months

- Integration of Unicode Detection Toolkit with ACP
- Prototype
- Plug-and-play annotators
- Expanded participation in TREC (Chinese)