



URSA

Unicode Retrieval System Architecture

Unicode Retrieval System Architecture

(URSA)

TIPSTER Phase III

Mark Davis & Bill Ogden

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

CRL COMPUTING
RESEARCH
LABORATORY





Unicode

- A coding system for most of the World's languages.
- Unifies common characters (e.g. punctuation and Japanese-Chinese ideograms) to single codepoints.
- Uses “composed” characters to represent accented and ligatured forms of characters.
- Has “equivalency” algorithms to determine if composed and non-composed characters are the same.
- Simplifies multilingual text processing.



Unicode Detection

- Indexes Unicode text strings and characters
- Converts other character sets to Unicode
- Normalizes text using equivalency algorithm
- Uses index terms from “linguistic” TIPSTER document annotators
- Has Boolean, Proximity and Natural Language queries



URSA Detection

Русский Язык

Ամերկահայ

მხედრული

대한민국

中国

日本語

ไทย

ລາວ

- Indexes TIPSTER document collections
- Uses TIPSTER annotations to get index terms
- Uses TIPSTER queries to retrieve documents
- Interfaces with URSA display technology for document display and query specification
- Includes conversion to-and-from Unicode for 80 different character encodings.



URSA Display Technology

- Based on CRL's MUTT (Multilingual Unicode Text Toolkit) to provide display and editing capabilities in over 40 languages.
- **Scripts - CHINESE JAPANESE KOREAN HEBREW ARABIC GREEK ETHIOPIC LATIN CYRILLIC ARMENIAN GEORGIAN LAO THAI**
- **Languages - AMHARIC, ARABIC, ARMENIAN, GEORGIAN, GREEK, HEBREW, JAPANESE, KOREAN, PASHTO, RUSSIAN, SERBO-CROAT, SIMPLIFIED & TRADITIONAL CHINESE, THAI, AND ALL LATIN BASED LANGUAGES**



URSA User Interface Technology

- Designed based on analyst's needs
- TUIT - TIPSTER User Interface Toolkit
COLLECTION OF GUI MODULES FOR VIEWING AND MANAGING TIPSTER OBJECT (DOCUMENTS & ANNOTATIONS.)
- Annotation plug-ins.
CONVERTERS FROM NLP TAGGERS TO TIPSTER ANNOTATIONS.



Example TIPSTER Applications

- **Oleada** - Multilingual environment for language teachers and learners
- **Tabula Rasa** - MUC style text extraction GUI builder
- **QUILT** - Cross-Language Information Retrieval system
- **Corelli** - Rapid development environment for machine translation



QUILT (Query User Interface with Light Translations):

- English queries for Spanish documents.
- Query translation based on tag-based bilingual dictionary and corpus-based disambiguation.
- Vector-based retrieval engine.
- Gloss-translations of retrieved documents.
- Scan alternative translation terminology for retrieved document.
- Web-based components.



Query Translation

● **English Query**

How has the threat of swine fever affected international trade?

● **Tagged English Query**

VB_has NN_threat NN_swine
NN_fever VB_affected
JJ_international NN_trade

● **Stemmed Query**

NN_threat NN_swine NN_fever
VB_affect JJ_intern
NN_trade

● **Dictionary Lookup**

NN_threat achor|amag|amenaz|bravat|conmin|disfuerz|espant|nublar|pe
NN_swine canall|cochin|galduf|jet|malaj|mam|marran|papal|perr|puer
NN_fever calentur|chuch|fiebr|pasm
NN_intern intern
NN_trade comerc|contrat|negoc|ofic|sindicat|tráfag|tráfic|trapich

● **Corpus Disambiguated Query**

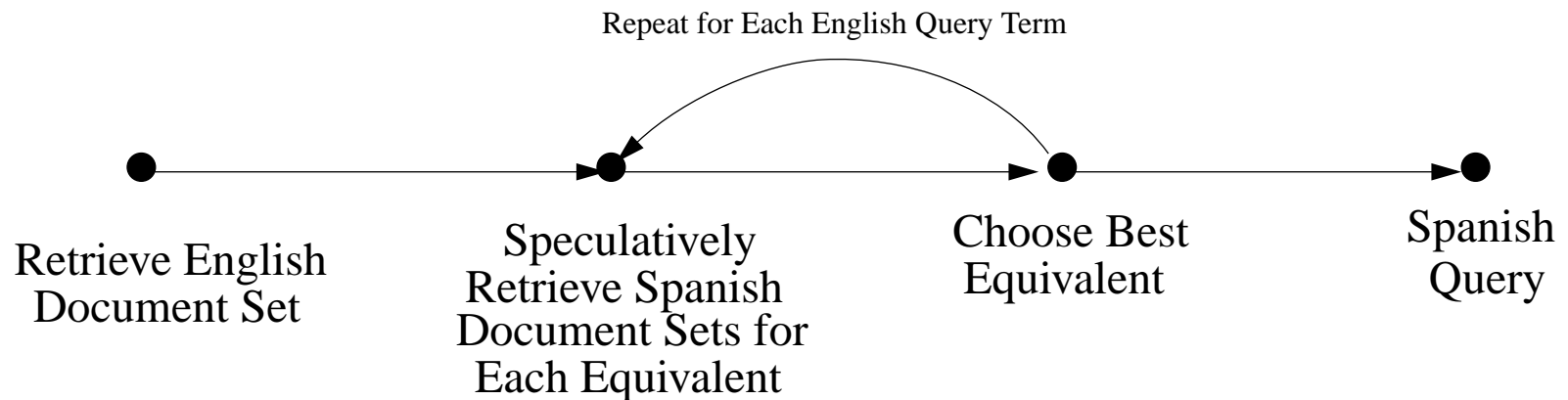
amenaz perr fiebr afect intern comerc

● **Spanish Retrieval**



Corpus Disambiguation

- Parallel English-Spanish aligned corpus (UN 1991).
- Around 100,000 alignment pairs.
- Alignments at the sentence or sentence-pair level.





Performance

<i>Method</i>	PR (NI)	% of MONO
<i>MONO</i>	0.2895	—
<i>QUILT</i>	0.2127	73.5
<i>POS</i>	0.1949	67.3
<i>ALL</i>	0.1422	49.1
<i>CORP</i>	0.1153	39.8



Conclusions and Ongoing Work

- Evaluation Java for high-performance multilingual text retrieval
- Improved document translation strategies.
- Evaluation of CLTR user needs to improve upon GUI and model.
- Unicode/TIPSTER compliant basic retrieval engine.