

# Foreign Language Exploitation Tools and Experiments

Dr. David S. Day

781-271-2854 • [day@mitre.org](mailto:day@mitre.org)

Army-Contract MOIE



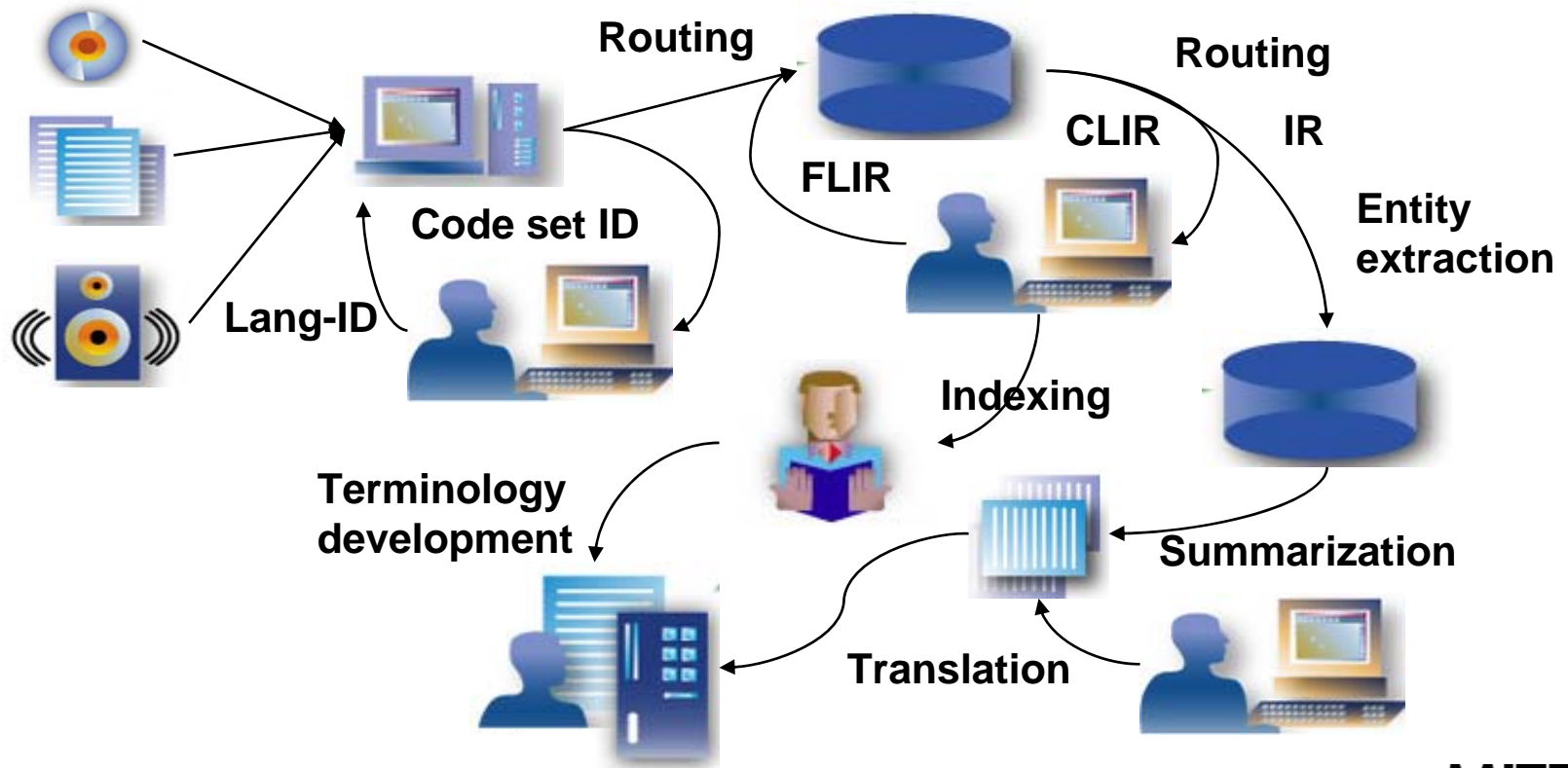
**MITRE**  
Technology  
Program

# Problem

- **There is a huge amount of critical foreign language material that fails to be exploited due to insufficient numbers of skilled translators and language analysts.**
- **Purely automated approaches alone are not able to capture the meaning and significance of these foreign language materials.**
- **Can emerging state-of-the-art tools and resources aid human translators/analysts to enhance their productivity or performance?**

# Background

- The foreign language exploitation process depends upon the diverse roles and skills provided by human linguists and translators



# Objective

- **Demonstrate and measure the impact of state-of-the-art computational aids and resources on linguist productivity and performance**
- **Assess as a function of various factors**
  - **Task type (summary, full translation, document triage/routing, reading comprehension, ...)**
  - **Tools/resources available to analyst**
  - **Time constraints placed on linguist**
  - **Genre and other features of source materials**
  - **Linguist skill level**

# Activities

- **Fully instrumented the C/Flex foreign language exploitation experiment platform**
  - Can be invoked on documents directly from Clipper
- **Integrating commercial and research processing components and resources in three languages (Chinese, Arabic, Spanish)**
  - MT (Language Weaver, LEC, Systran); transliteration (Basis, Queens College); dictionaries, data and lexicons (LKS/Wordscape, LDC); translation memory, NLP (MITRE)
- **Reaching out to government linguists to participate in realistic experiments**
  - Available on OSIS, MITRE, and external networks

# Highlight



C/Flex Log File (simulated):

```
<ImportEvent file="xinhua-2005-10-14-1503.htm"
  TimeStamp= "2006-01-25-14-21-22-35" ID="422"/>
<LookupEvent resource="Adso-CE-Dict" string="      "
  TimeStamp="2006-01-25-14-21-15-31" ID="423"
  result="high-speed"/>
<LookupEvent resource="LangWeav" string="      "
  TimeStamp="2006-01-25-14-21-18-42" ID="532" ... />
<LookupEvent resource="CE-TM" string="      "
  TimeStamp= "2006-01-25-14-21-22-35" ... /> ...
<MTpaneEvent MTID="LangWeav" type="copy" .../>
...
<HTransPaneEvent type="paste" text="Many people ..."/>
...
```

- **Various hypotheses can be explored empirically based on fine-grained system logging:** Can summary productivity be increased without any measurable impact on quality? ... Do junior translators rely more on MT output than examining multiple dictionary entries? ... What tasks are least improved when current tools and resources are introduced? ...

# Demonstration: CalliFlex

Source text, pre-segmented, annotated for named entities

Translated text from multiple machine translation engines

The screenshot displays the CalliFlex software interface. At the top, a window titled "Callisto - ss-n-22-08-2005Se13.use-import\*" shows source text in Chinese, pre-segmented and annotated with various tags. A sidebar on the left contains a "Named Entities" histogram with categories like PERSON, ORGANIZATION, GPE, LOCATION, FACILITY, DATE, TIME, MONEY, and PERCENT. Below this is an "Integrated chat" window. The main area shows two machine translations: "MT: LEC" and "MT: LanguageWeaver". A "Translation Memory" panel on the right displays search results for the Chinese phrase "这种看法是". At the bottom, there are panels for "Human translation or summary", "Multilingual dictionaries and other searchable lexicons", and "Translation Memory".

Human translation or summary

Named entity histogram

Integrated chat

Multilingual dictionaries and other searchable lexicons

Translation Memory

MITRE

© 2006, The MITRE Corporation

# Impacts

- **Provide concrete empirical analyses that reveal whether, and in what ways, tools and resources can enhance human exploitation productivity**
  - Identify targets of opportunity for government focus, vendor investment, academic research
- **Demonstrate richly articulated user interface of tightly integrated text processing tools**
  - User-customizable multi-pane GUI for arranging and interacting with multiple components
  - Service-oriented architecture encourages rapid integration of new capabilities (local or networked)

# Future Plans

- Execute plans for detailed empirical studies
- Expand user-editable resources to include
  - User-vetted, search-engine supported, normalized entity tracking (specific people, organizations, places)
  - Standardized and alternative transliterations for individuals
- Fine-grained tracking of copyrighted data and licensed tools
- Mixed-initiative annotation of documents for improved triage and routing

