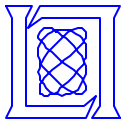


CTIDES and Its Applications to US-Korea Joint Digital Libraries Initiatives

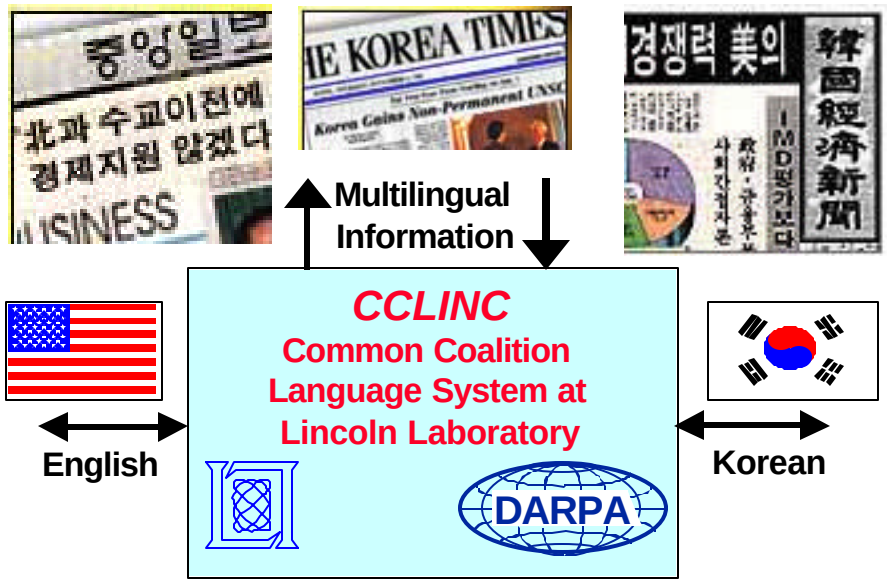
Presentation at US-Korea Joint Workshop on Digital Libraries

Young-Suk Lee

August 10, 2000

- **CTIDES: Coalition Translingual Information Detection, Extraction and Summarization**
- **CCLINC Translingual Information Technology Overview**
  - **English-Korean Text Translation**
  - **Translingual Information Detection, Extraction and Summarization**
- **Position Statement and Technical Challenges**

Coalition Translingual Information Detection, Extraction & Summarization



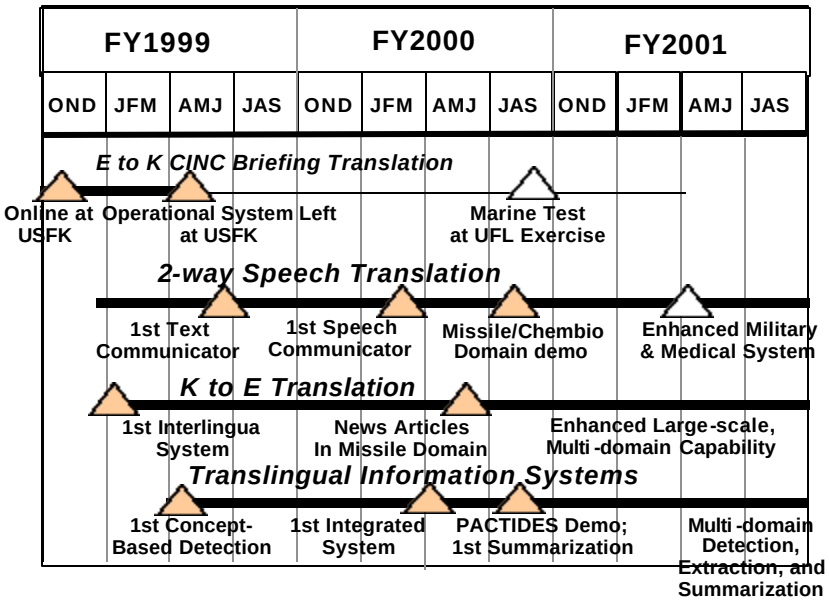
New Ideas

- Two-way, interactive English/Korean speech translation
- First large-scale interlingua-based Korean-to-English translation
- Understanding-based, domain-independent, translingual information extraction and summarization
- Understanding-based approach for high precision translingual information retrieval
- Integrated human/human and human/machine interactive system architecture

Impact

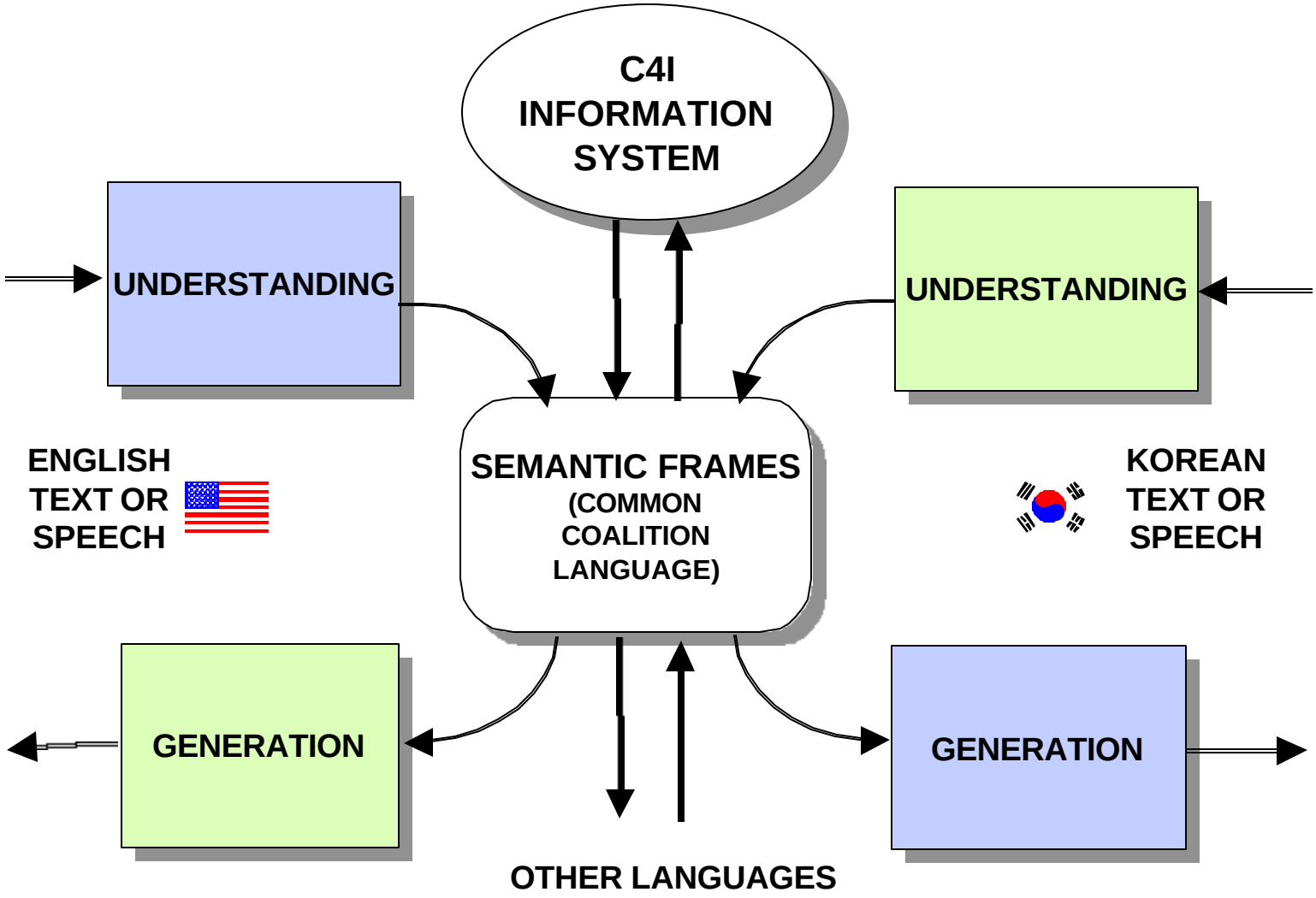
- Enhanced real-time communication with coalition allies via interactive speech translation
- Rapid translation of command briefings and reports for coalition collaboration
- Strategic situation awareness via automated translingual access to worlwide, multilingual sources of C4I information
- Powerful Force Multiplier via enhanced effectiveness of coalition C4I

Schedule



PIs: Young-Suk Lee and Clifford Weinstein

CCLINC Translingual Information System Structure

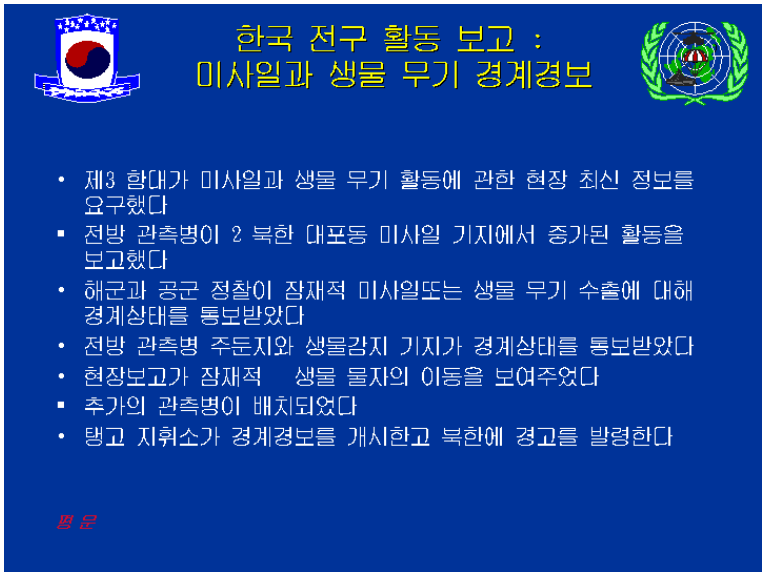


CCLINC = Common Coalition Language System at Lincoln Laboratory

English-to-Korean Text Translation

(09/01/95 --- 04/30/00)

- System trained on about 8,000 sentences from CINC (Commander-in-Chief) powerpoint briefings
- About 70% accuracy on test data from the same domain
- About 80% automated acquisition of grammars/lexicons
- About 35K distinct entries in the bilingual lexicon
- Technology transfer to MFP (marine forces pacific) in progress



Korean-to-English Document Translation

(09/01/99 – Present)

- Trained on about 6K sentences from Korean newspaper articles (22 words/sentence)
- 80% automated acquisition of lexicons/grammars
- About 20K entries in the bilingual lexicon
- About 60% accuracy on training data, 45% accuracy on test data
- Quality improvement in progress

미국 국방부는 24일 발표한 화학.생물무기에 관한 보고서에서 북한이 지난 89년이후 신경가스 등을 대량 생산할 수 있는 능력을 갖고 있으며, 현재 상당량의 화학무기를 보유하고 있다고 밝혔다.

교도(공동)통신의 워싱턴발 보도에 따르면 보고서는 또 북한이 생물무기에 대해서도 지난 60년대 이후 개발노력을 계속해 오고 있다면서, 핵.미사일과 더불어 북한의 대량파괴무기 개발 위협이 고조되고 있다고 지적했다.

보고서는 북한의 화학무기 계획의 수준이 매우 높다면서 신경, 피부, 호흡기, 혈액 등에 장애를 일으키는 무기를 대량 생산할 수 있어 한국과 교전시의 사용을 상정하고 있는 것으로 분석했다.

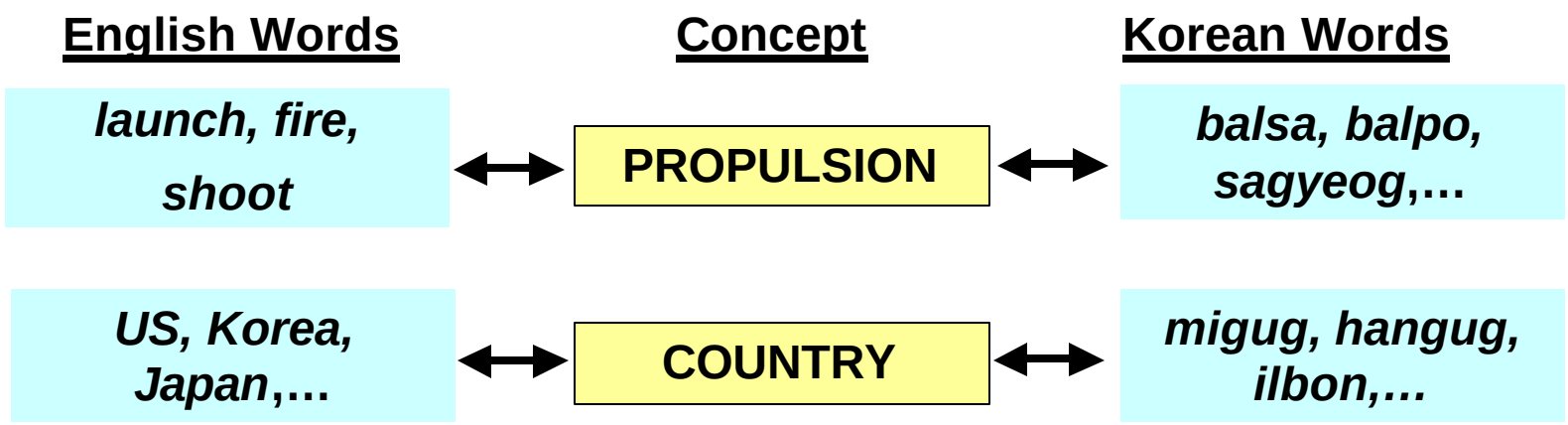
North Korea retain big amount's chemical weapon

As for US Department of Defense announce 24 date. North Korea at report can mass produce nerve gas regarding chemical and biological weapon after last 89 year. it have capability and it declared it retain at the present time big amount's chemical weapon according to Gyoto communication company's Washington edition walkway. As for report North Korea continue development effort again regarding biological weapon after last 60 years and North Korea's mass destruction weapon development threat increase with nuclear and missile, it pointed out. As for report chemical weapon North Korea's level be high very and induce difficulty in nerve and skin and respiratory organs and blood. Can mass produce weapon. It analyzed it presuppose Korea and engagement time's use, as for North Korea it conduct discipline extensively again regarding protection mask and protection clothes wearing elimination method, report declared.

Translingual Information Detection, Extraction and Summarization

- **Mode of Operation**
  - Query in English
  - Concept-Based search on English and Korean documents
  - Summarization and translation of relevant documents to English
- **General purpose English concept lexicon with 300K entries**
- **Machine learning technique for**
  - Predicting the category of new words
  - Disambiguating the category of ambiguous words

Both Query Keywords and Database are Translated into Concepts for Search

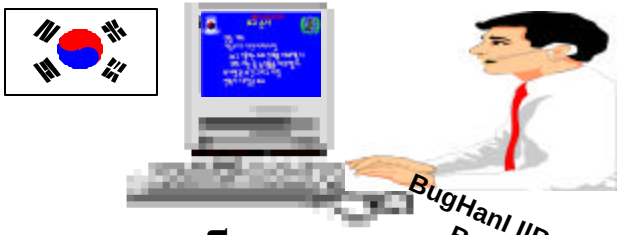


Example: Translingual Information Retrieval and Translation

Query: Which country has chemical biological weapons?

TITLE	LANGUAGE	TRANSLATION	SUMMARY
North Korea retain big amount's chemical weapon		T	
chemical weapon attack time nuclear weapon response		T	
retain North Korea chemical weapon ton 5000		T	
recommend forced US House of Representative subcommittee and US force anthracnose vaccine vaccination abeyance		T	
strengthen USFK and chembio–war preparation		T	
3 lethal gas chemical–warfare capability world above		T	
ISRAELIS ANXIOUS OVER POSSIBLE IRAQI ATTACK			SUMMARY
modif Seoul Seoul defense goal oplans 5027		T	
North Korea's chemical–warfare preparation chemical–biological–and–radiological protection headquarter establishment		T	

Integrated CCLINC Translingual Information System



BugHanI IlBonE Ddo MiSalIEuI
BalSaHan Deus HaDa

2-way Speech Translation

CCLINC SERVER

What is the north korean movement in the front line?

What is the status of nk missile launch against japan?

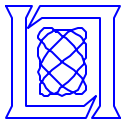
Translation

It seems that North Korea launch a missile again
After North Korea launched a Daipodong missile last month, NK is perceived to proceed to an additional test launch. Korea, US and Japan enter into an alert state, and prepare for a joint response policy. Korea estimates that the additional launch will be on 09/05. Japan estimates that NK's missile range is short. US information says that there is no sign of launch yet.

Position Statement

3 steps towards getting results by 2003

- **Step 1: Determination of content & collection**
  - **Work with pre-existing online resources (i.e. trying to build a new online resource should be a longer term goal)**
  - **Technical reports, research papers, dissertations**
- **Step 2: Resource sharing in the areas of interest**
  - **Bilingual lexicons**
  - **Most of other resources can be easily derived from bilingual lexicons**
- **Step 3: Application of currently available information technology (e.g. CCLINC) for machine translation and translingual information retrieval**
  - **Building a new technology from scratch is not the way to go**



Surmountable Technical Challenges

- **Machine Translation**
  - **Technology**
    - **High quality translation output in various subject areas**
  - **User Interface**
    - **Preservation of input and output format**
  - **System**
    - **Character set and code conversion**
- **Crosslingual Information Retrieval**
  - **Technology**
    - **accurate term translation in various subject areas**
    - **translation of unknown query terms**
  - **System**
    - **searching distributed collections**
  - **User Interface**
    - **can provide an intermediate solutions to the technology bottleneck via display of bilingual lexicon, synonym list, etc.**