Towards the Cross-Language Technology

Virach Sornlertlamvanich
Information Research and Development Division
NECTEC, Thailand
virach@nectec.or.th

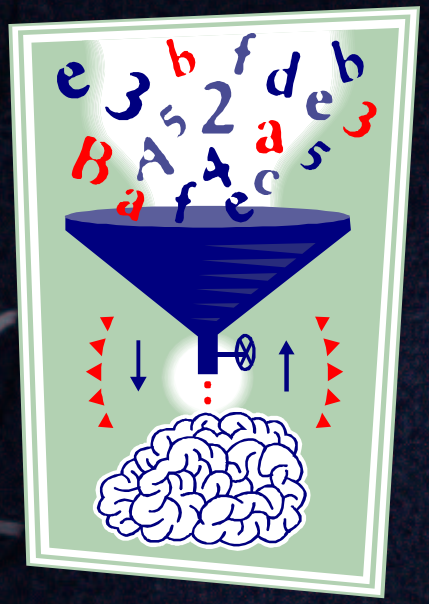
APAN 2003 Conference in Fukuoka, 21-24 Jan 2003

- ParSit */parse-it/*
  - ภาษิต (proverb)
  - English-Thai cross-language web nevigator.
  - www.suparsit.com

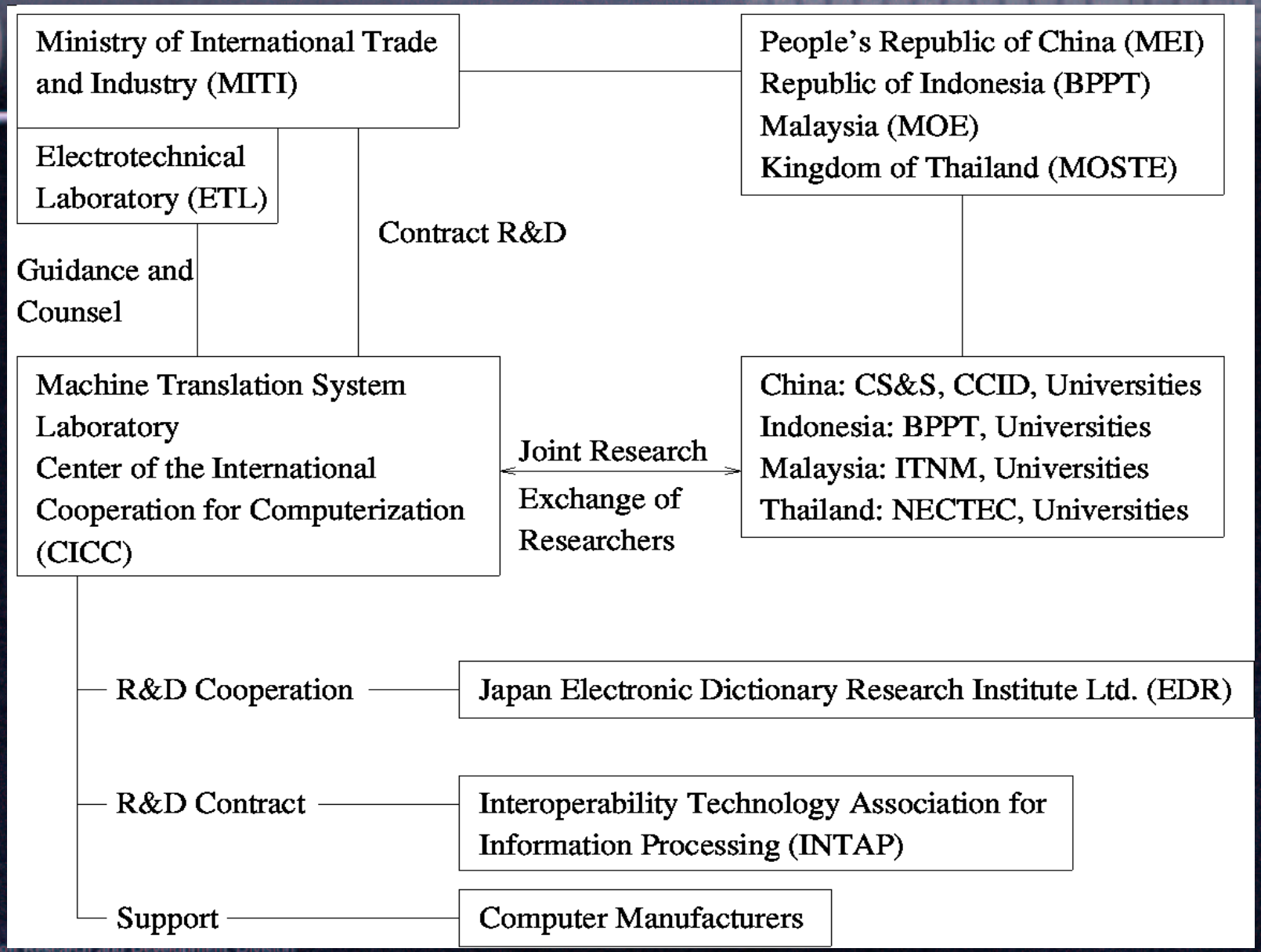
2 Big Machine Translation Projects

- ARIANE Project (1981-)
- Multilingual Machine Translation Project (1987-94) C, I, M, J, T

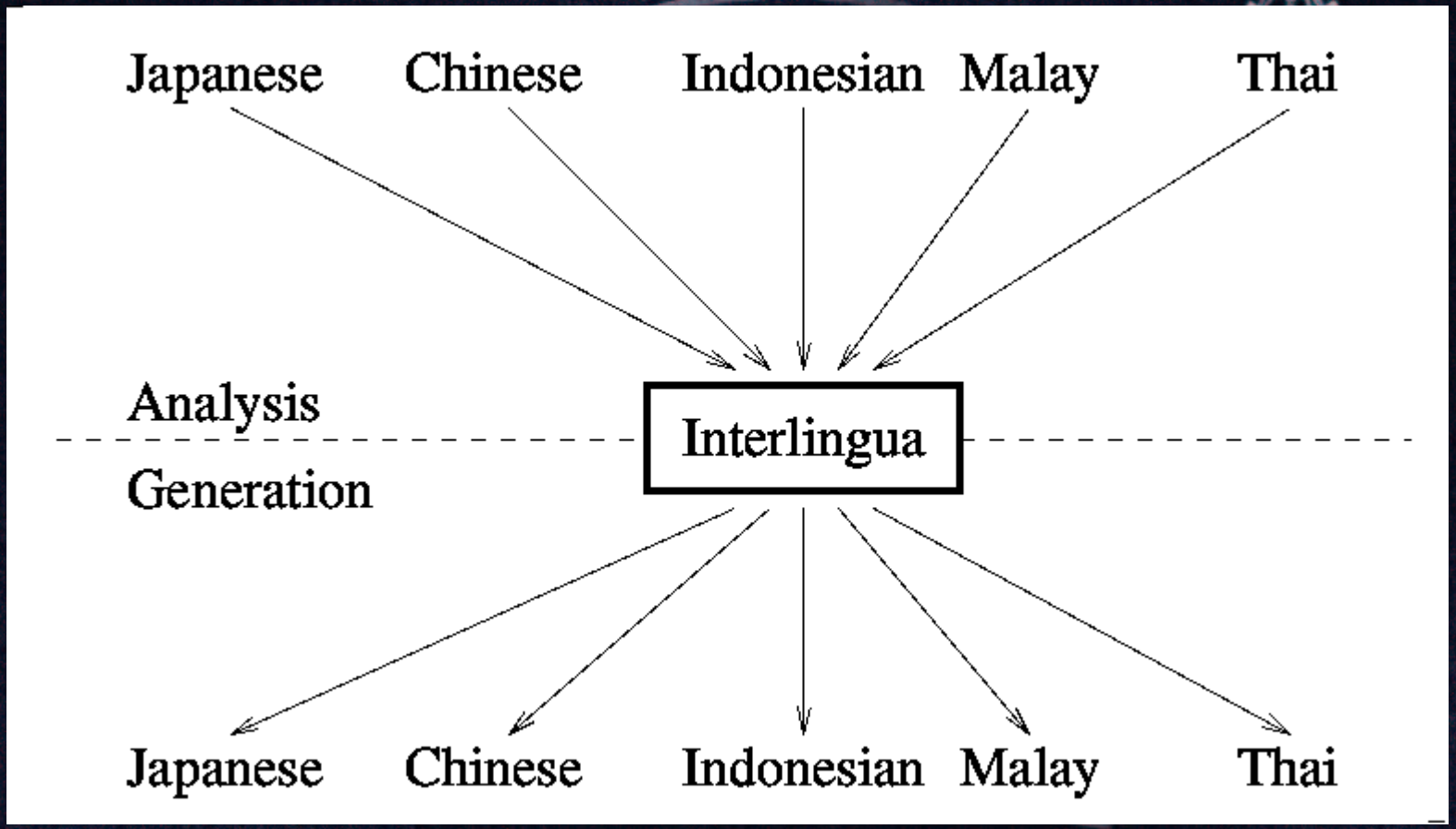
Learn from experienced countries



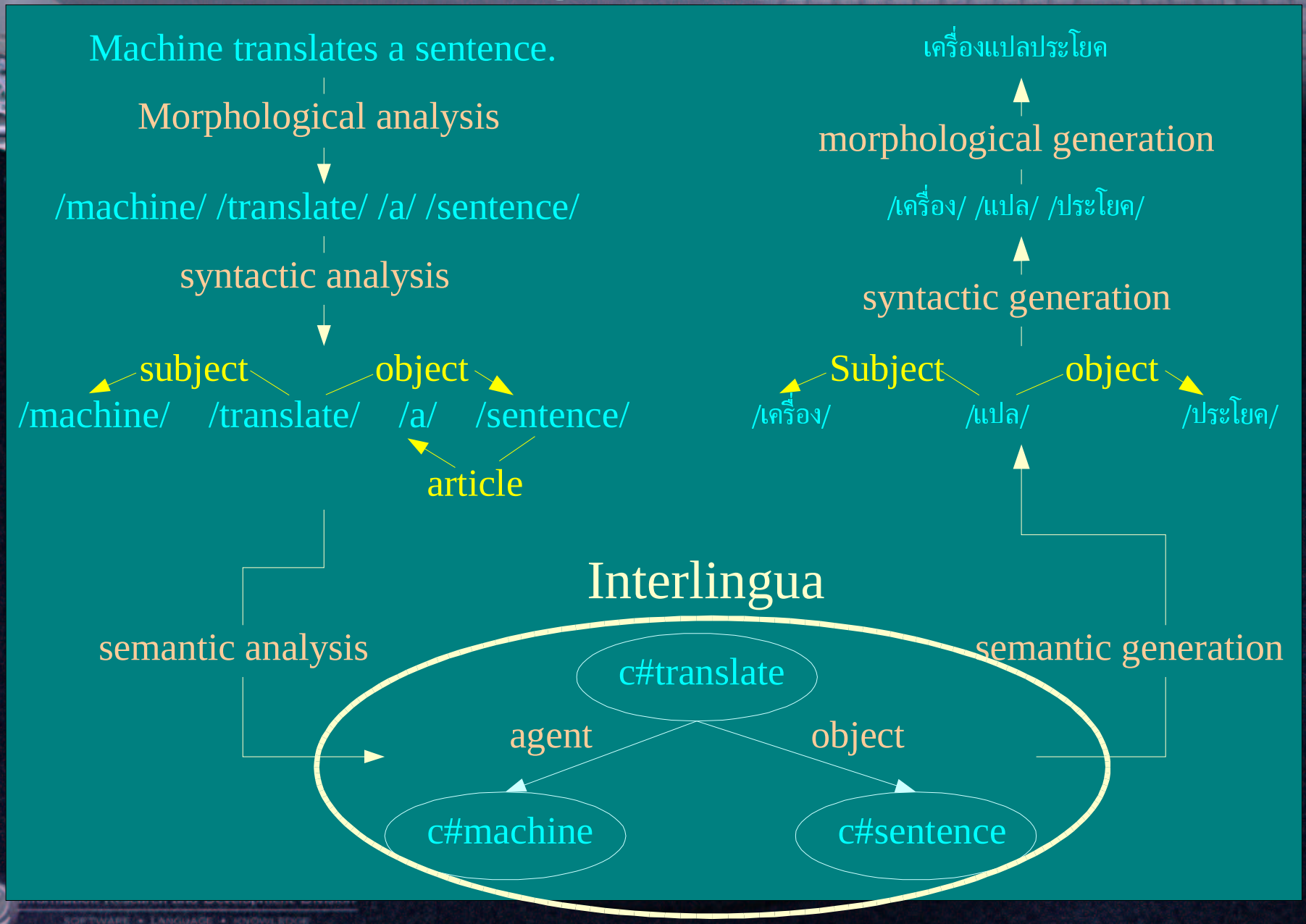
Multilingual Machine Translation Project



Interlingual Approach



Interlingual Approach

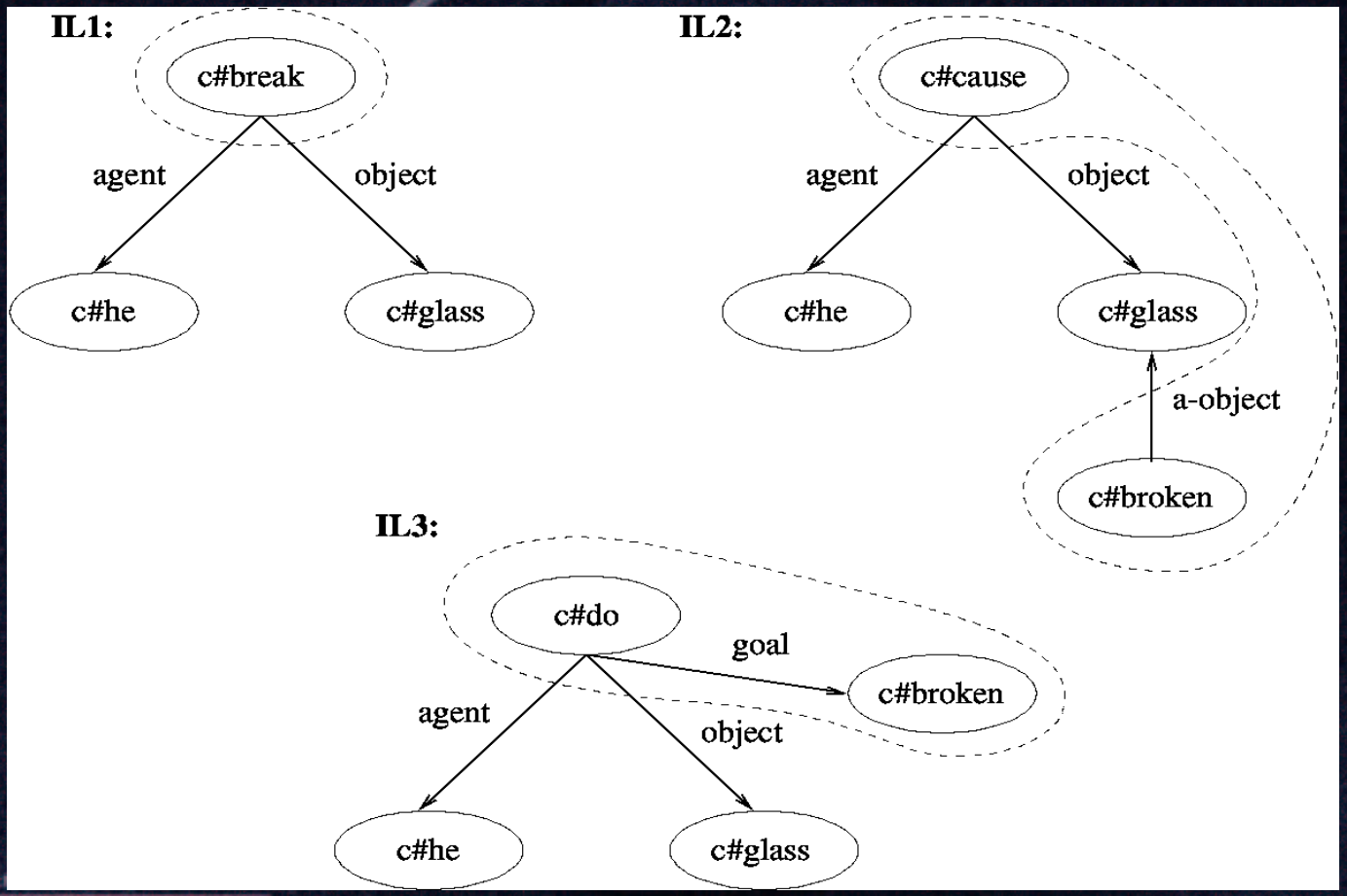


Difficulites in MT

- Grammatical issues
  - Inflection
  - Article
  - Grammatical case marker
  - Word and sentence boundary, Punctuation
- Semantic issues
  - Definitions of concept, set of semantic case, ontology
  - Concept decomposition, concept divergency, concept granularity

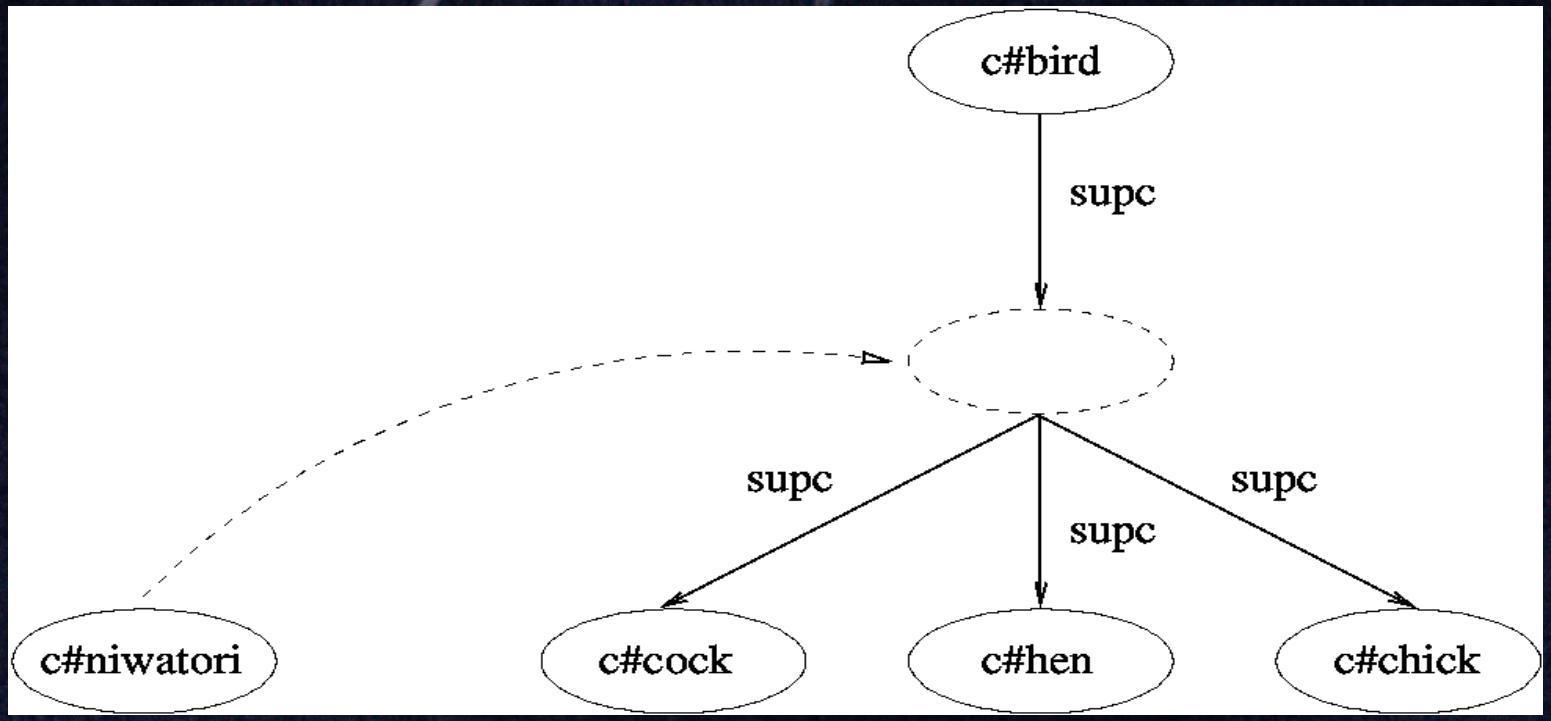
Concept decomposition

He breaks the glass. [He causes the glass broken.]



Concept divergency

A concept of Language A is corresponding to a meta-concept in Language B.

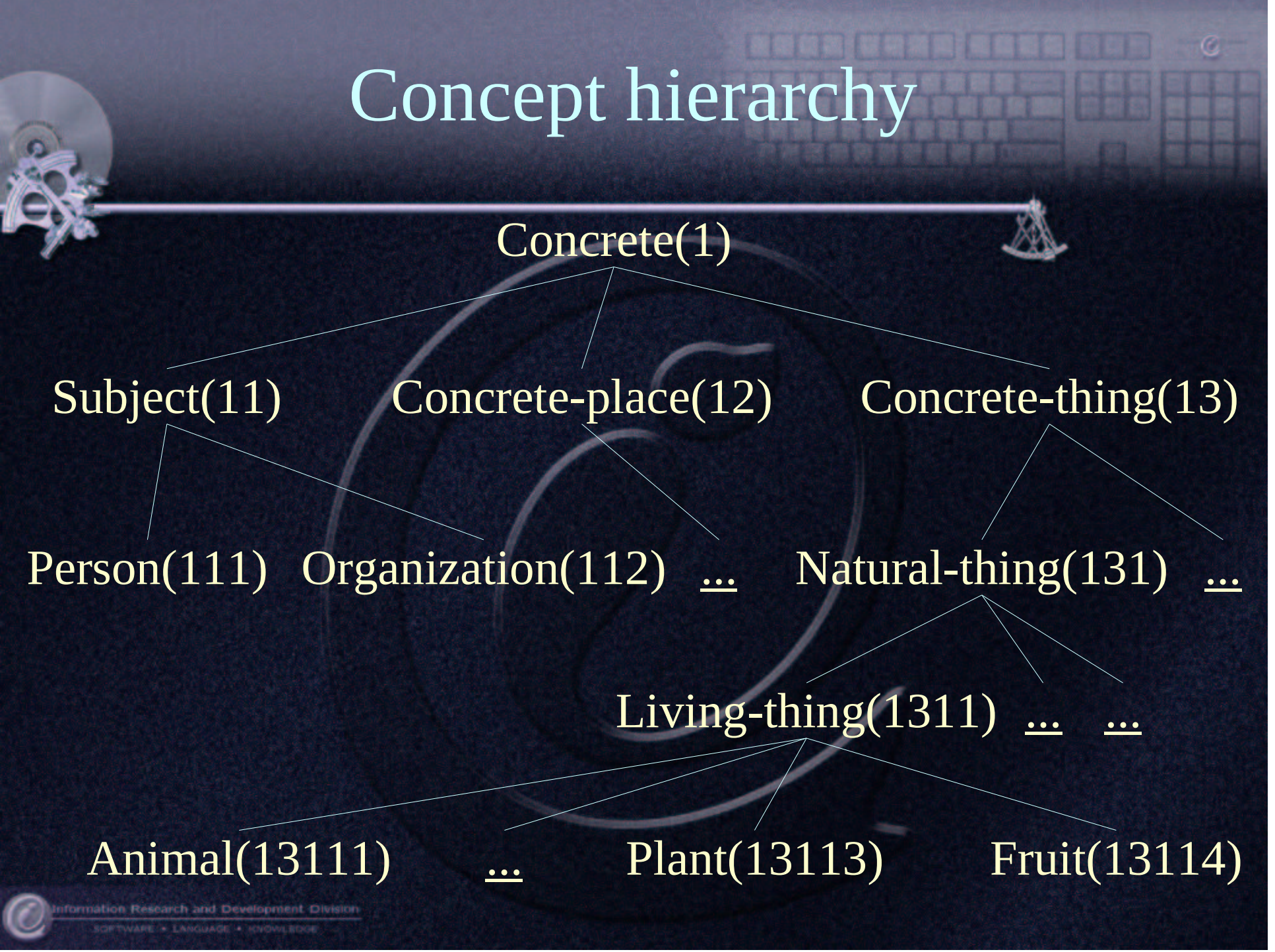


Concept granularity

A concept of Language A is corresponding to a multiple concepts in Language B.

Japanese	English
sakura (tree) (flower)	cherry tree cherry blossom
sakuranbo (fruit)	cherry (fruit) (tree)

Concept hierarchy



Ontologies

- EDR
  - Approach: Word description as employed in dictionaries
  - Problem: Ambiguities and incomputability
- Wordnet
  - Approach: Synonym set and simple semantic relations to other words
  - Problem: Ambiguities
- UW
  - Approach: Headwords and semantic restrictions
  - Advantage: Computability and no ambiguity

Ontologies

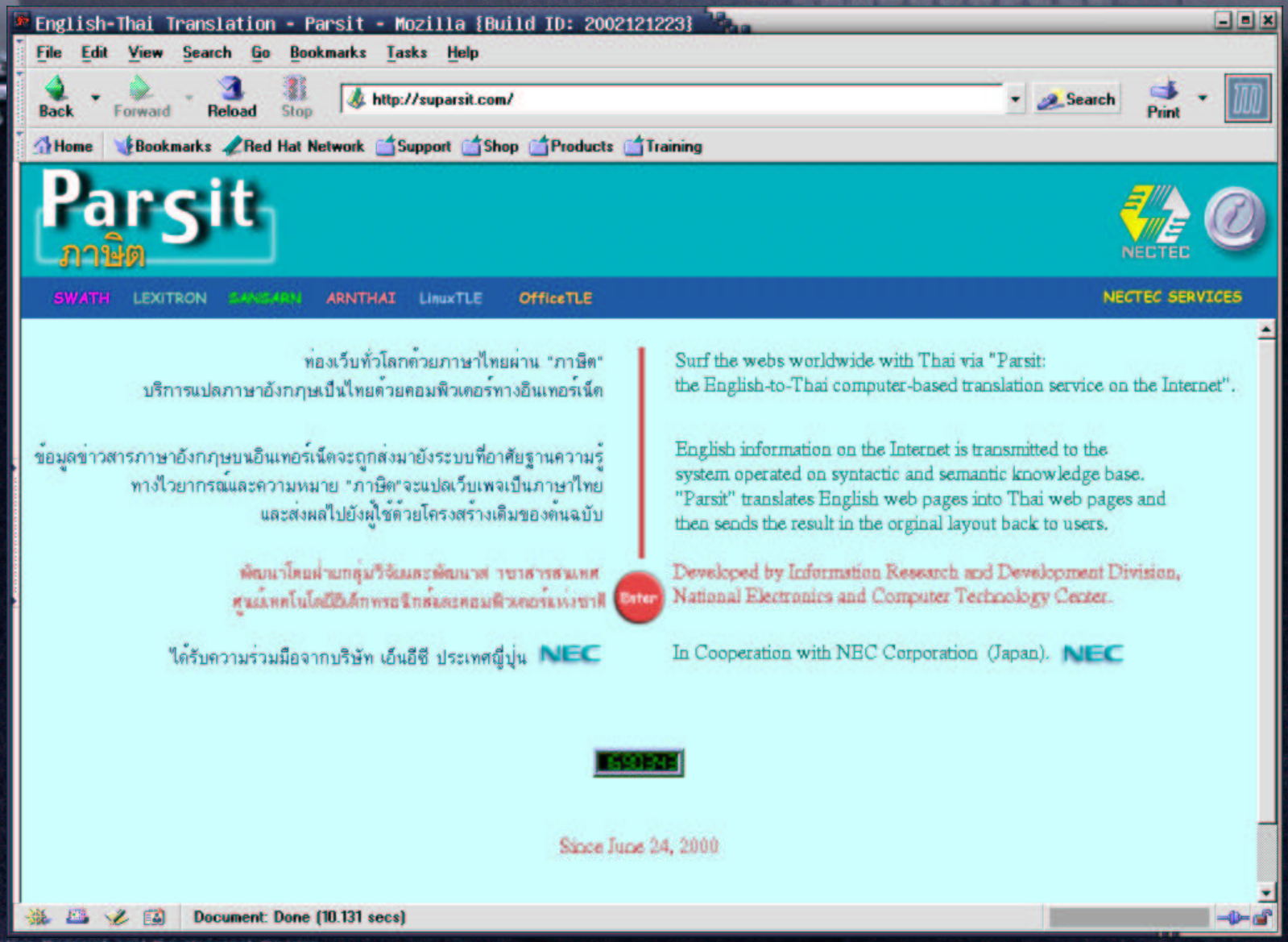
Representation of concept 'tired' in different schemes

EDR	Wordnet 1.5	UW
- having or displaying a need for rest - having lost of interest - lack of imagination	- A1: tired (vs. rested) - A2: bromidic, commonplace, hackneyed, … - V1: tire, pall, grow weary, fatigue - V2: tire, wear upon, fag out - V3: run down, exhaust, sap, … - V4: bore, tire, ...	- tired - tired(icl>physical) - tired(icl>mental)

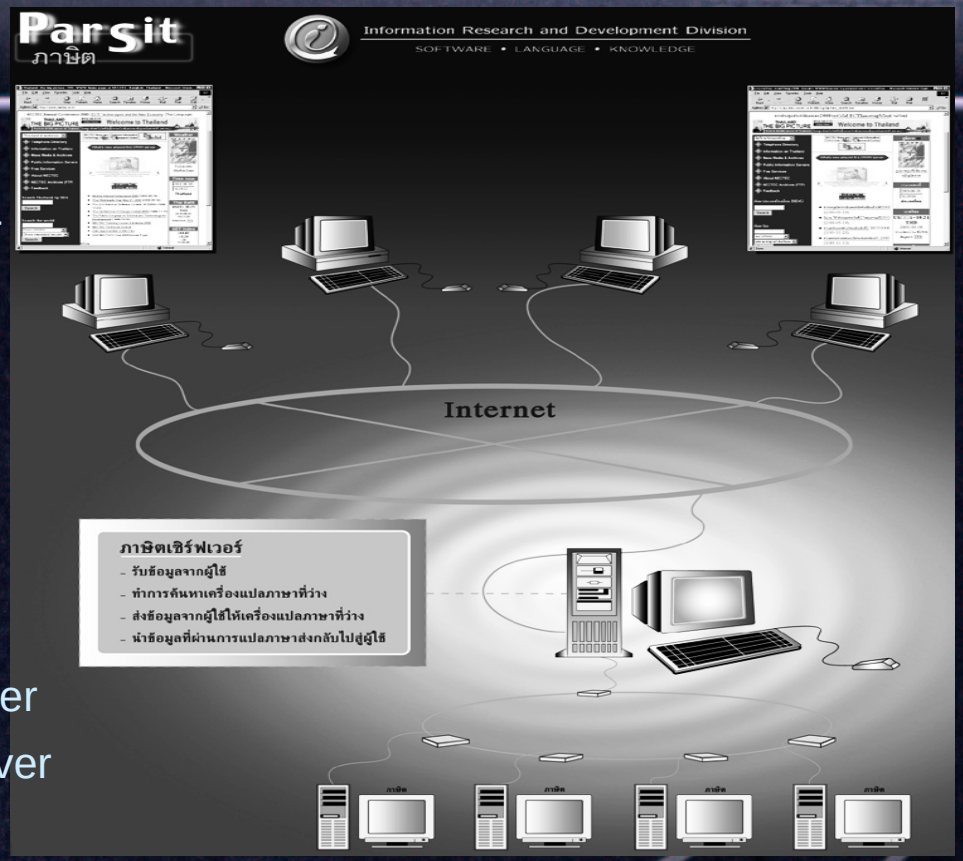
Design of Ontology

- Computational concept
- Concept insertion/deletion; composition/decomposition
- Expressive concept

www.suparsit.com



ParSit Service Process



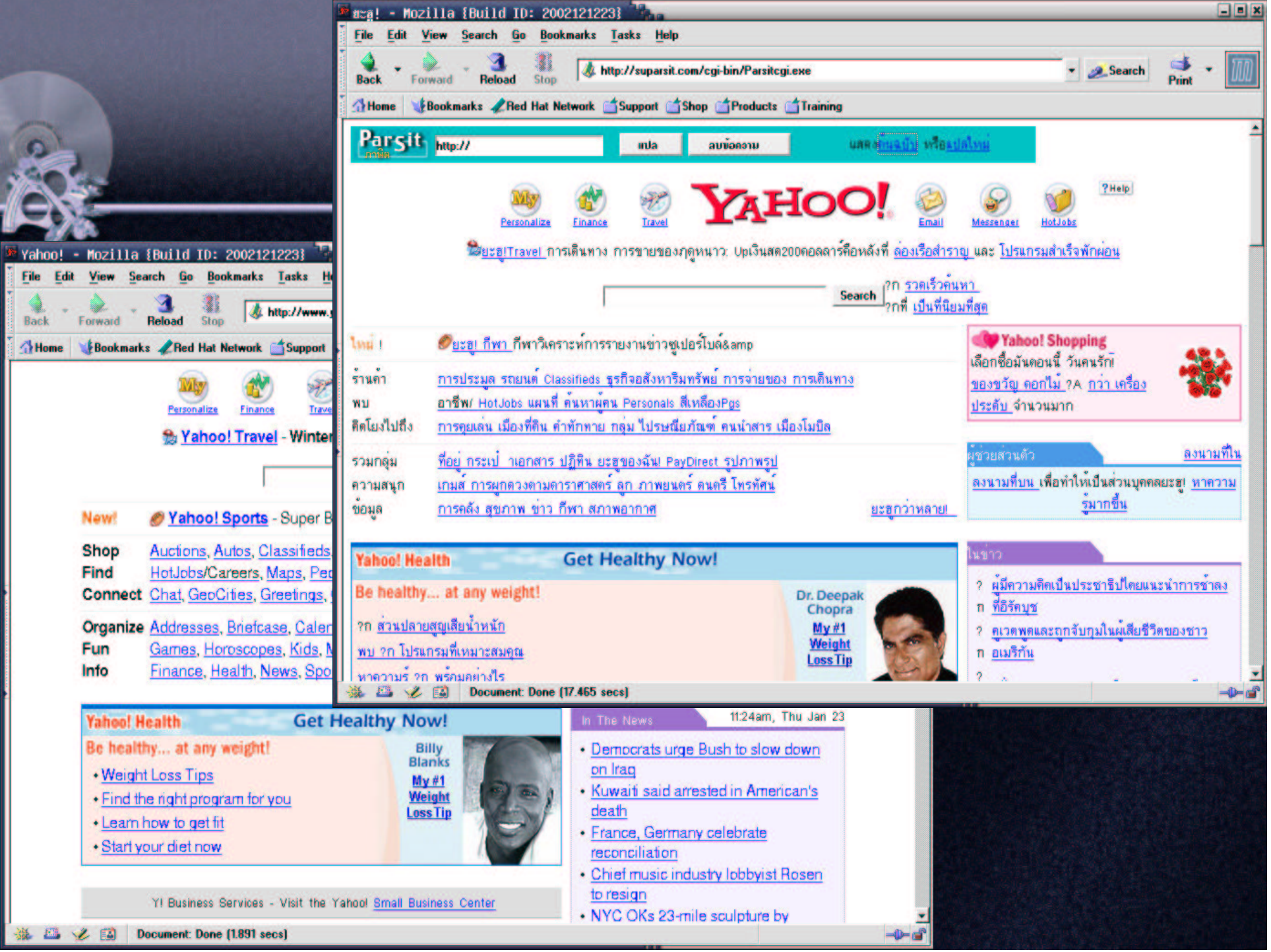
1 User send URL/TEXT to ParSit server

2 ParSit Server
- receive data from user
- send data to MT server

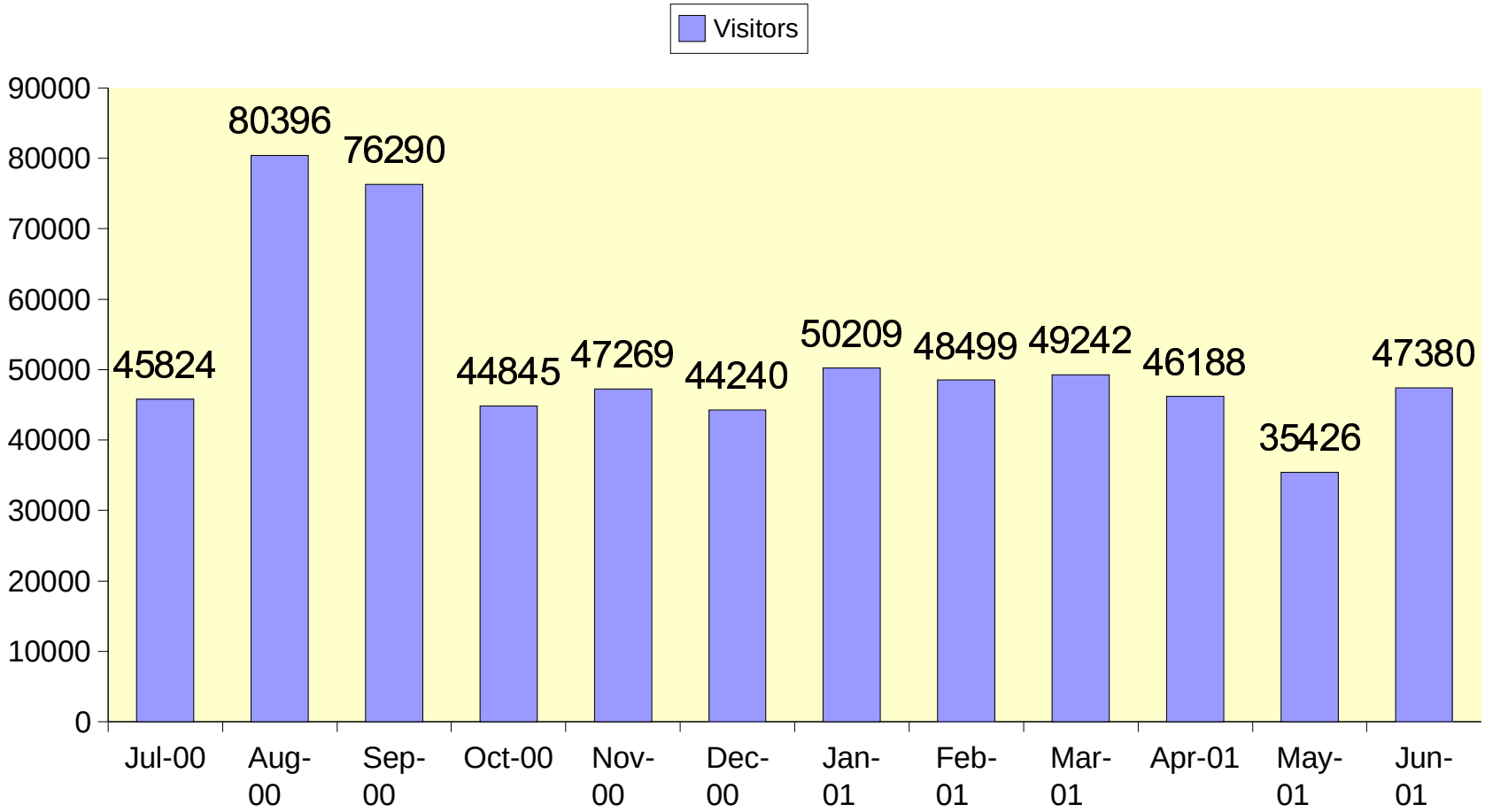
3 MT Server (E-T)

4 ParSit Server
- receive result from MT
- send result to User

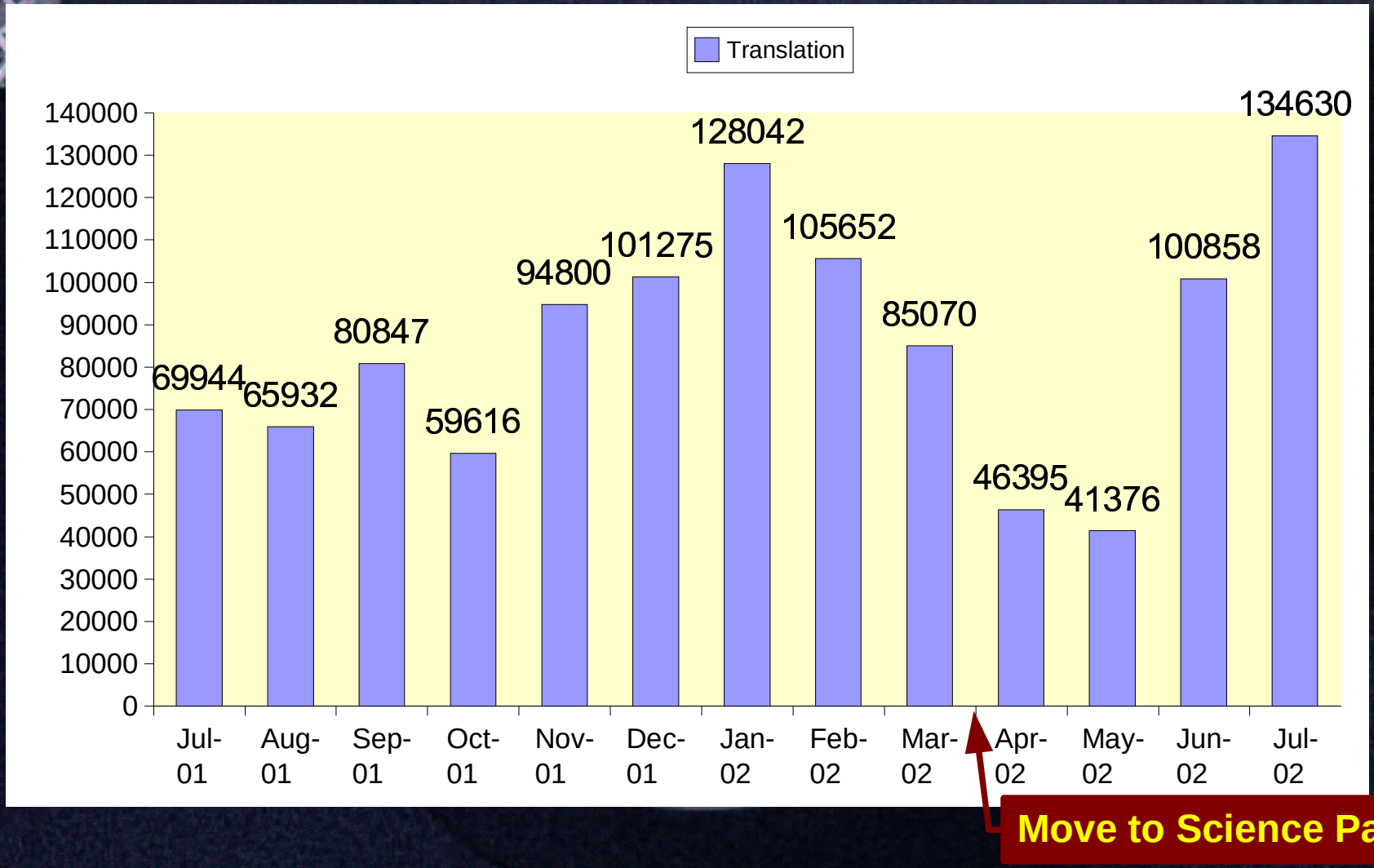
5 User receive Thai WP from server


Yahoo! - Mozilla {Build ID: 2002121223}
File Edit View Search Go Bookmarks Tasks Help
Back Forward Reload Stop
Home Bookmarks Red Hat Network Support
Personalize Finance
Yahoo! Travel - Winter
New! Yahoo! Sports - Super B
Shop Auctions, Autos, Classifieds
Find HotJobs/Careers, Maps, Ped
Connect Chat, GeoCities, Greetings,
Organize Addresses, Briefcase, Calen
Fun Games, Horoscopes, Kids, M
Info Finance, Health, News, Spo
Yahoo! Health Get Healthy Now!
Be healthy... at any weight!
Weight Loss Tips
Find the right program for you
Learn how to get fit
Start your diet now
Billy Blanks
My #1 Weight Loss Tip
Y! Business Services - Visit the Yahoo! Small Business Center
In The News 11:24am, Thu Jan 23
Democrats urge Bush to slow down on Iraq
Kuwaiti said arrested in American's death
France, Germany celebrate reconciliation
Chief music industry lobbyist Rosen to resign
NYC OKs 23-mile sculpture by
Document: Done (1.891 secs)
ยะฮู! - Mozilla {Build ID: 2002121223}
File Edit View Search Go Bookmarks Tasks Help
Back Forward Reload Stop
http://suparsit.com/cgi-bin/Parsitcgi.exe
Search Print
Home Bookmarks Red Hat Network Support Shop Products Training
Parsit
http://
แปล
ลบข้อความ
Personalize Finance Travel
YAHOO!
Email Messenger HotJobs
Help
ยะฮู!Travel การเดินทาง การขายของกูดูหนาว: Upเงินสด200ดอลลาร์คือหลังที่ ล่องเรือสำราญ และ โปรแกรมสำเร็จพักผ่อน
Search
รวดเร็วค้นหา
เป็นที่นิยมที่สุด
ใหม่ ! ยะฮู! กีฬา กีฬาวิเคราะห์การรายงานข่าวซูเปอร์โบล์&amp
ร้านค้า การประมูล รถยนต์ Classifieds ธุรกิจอสังหาริมทรัพย์ การจ่ายของ การเดินทาง
พบ อาชีพ/ HotJobs แผนที่ ค้นหาผู้คน Personals สีเหลืองPgs
คิดโยงไปถึง การคุยเล่น เมืองที่ดิน คำทักทาย กลุ่ม ไปรษณียภัณฑ์ คนนำสาร เมืองโมบิล
รวมกลุ่ม ที่อยู่ กระเป๋ าเอกสาร ปฏิทิน ยะฮูของฉัน! PayDirect รูปภาพรูป
ความสนุก เกมส์ การผูกดวงตามดาราศาสตร์ ลูก ภาพยนตร์ ดนตรี โทรทัศน์
ข้อมูล การคลัง สุขภาพ ข่าว กีฬา สภาพอากาศ
ยะฮูกว่าหลาย!
Yahoo! Shopping
เลือกซื้อมันตอนนี้ วันคนรัก!
ของขวัญ ดอกไม้ ?A กว่า เครื่อง
ประดับ จำนวนมาก
ผู้ช่วยส่วนตัว
ลงนามที่ใน
ลงนามที่บน เพื่อทำให้เป็นส่วนบุคคลยะฮู! หาความ
รู้มากขึ้น
Yahoo! Health Get Healthy Now!
Be healthy... at any weight!
ส่วนปลายสูญเสียน้ำหนัก
พบ ?ก โปรแกรมที่เหมาะสมคุณ
Dr. Deepak Chopra
My #1 Weight Loss Tip
ในข่าว
ผู้มีความคิดเป็นประชาธิปไตยแนะนำการช้าลง
ที่อิรัคบุช
คูเวตพูดและถูกจับกุมในผู้เสียชีวิตของชาว
อเมริกัน
Document: Done (17.465 secs)

Number of visitor/month



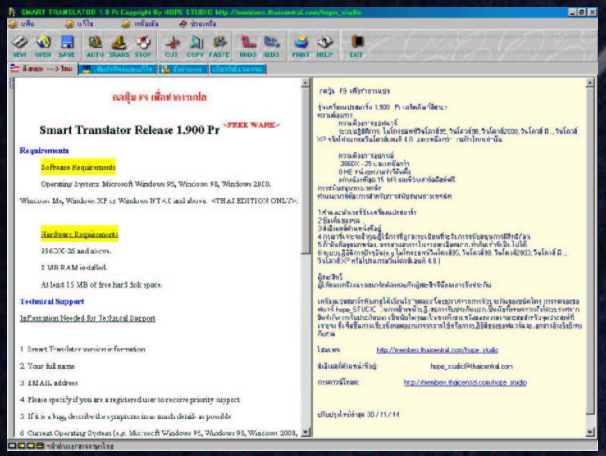
Number of unique session/month

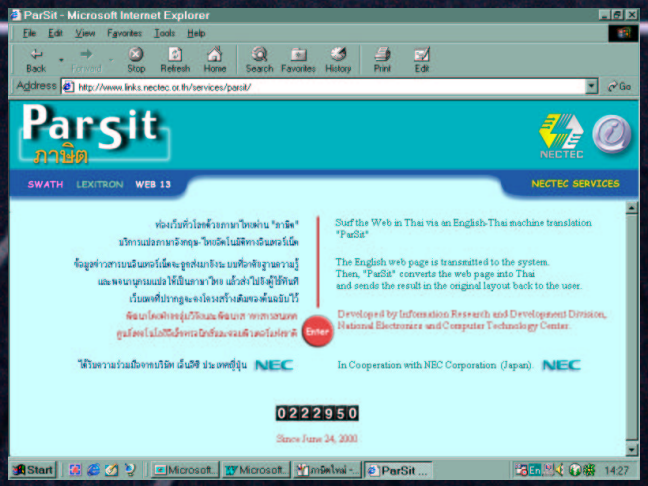


Available MT in Thailand

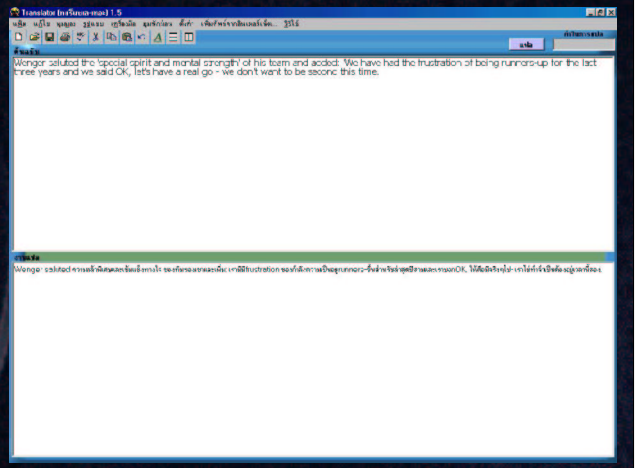
ParSit

SMART TRANSLATOR 1.9

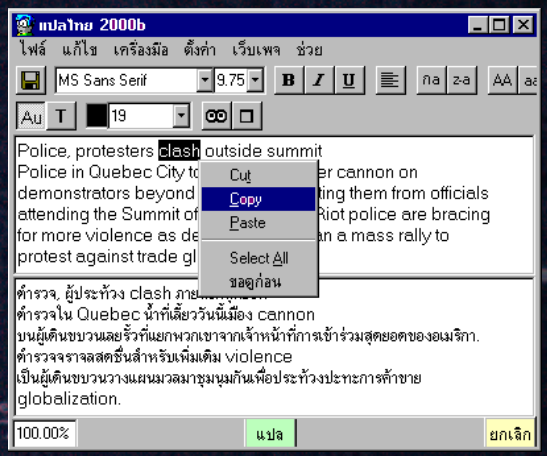


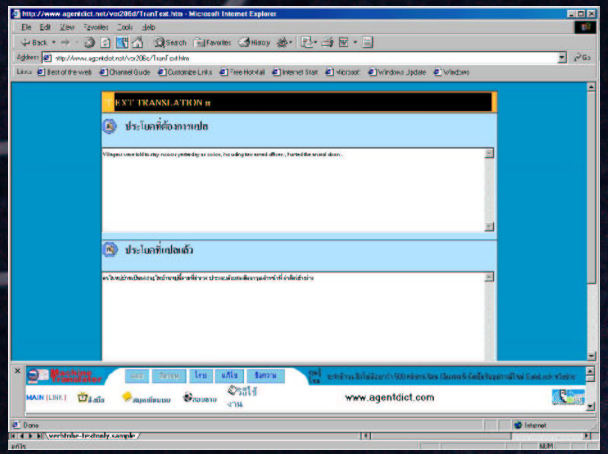


TRANSLATOR 1.5



PLAETHAI





AGENT DICT

Available MT in Thailand

- ParSit
  www.suparsit.com
- Smart Translator 1.9
- Translator 1.5
  www.the-kstudio.com
- PlaeThai 2000b
  www.palthai.com
- AgentDict Translation v.2.06
  www.dinosoft.co.th/software/AgentDict/AgentDict.htm

MT Systems in Thailand

	Parsit	Smart Translator	PlaeThai	Translator 1.5	AgentDict
Language	Eng-Thai	Eng-Thai	Eng-Thai	Eng-Thai	Eng-Thai
Lexicon	80,000	33,400	11,000	8,000	> 200,000
Type of document	-Web page -Text	-File (Text) -Text	-Web page -File (Text) -Text	-File (Text) -Text	-Web page -Text

Evaluation Corpus

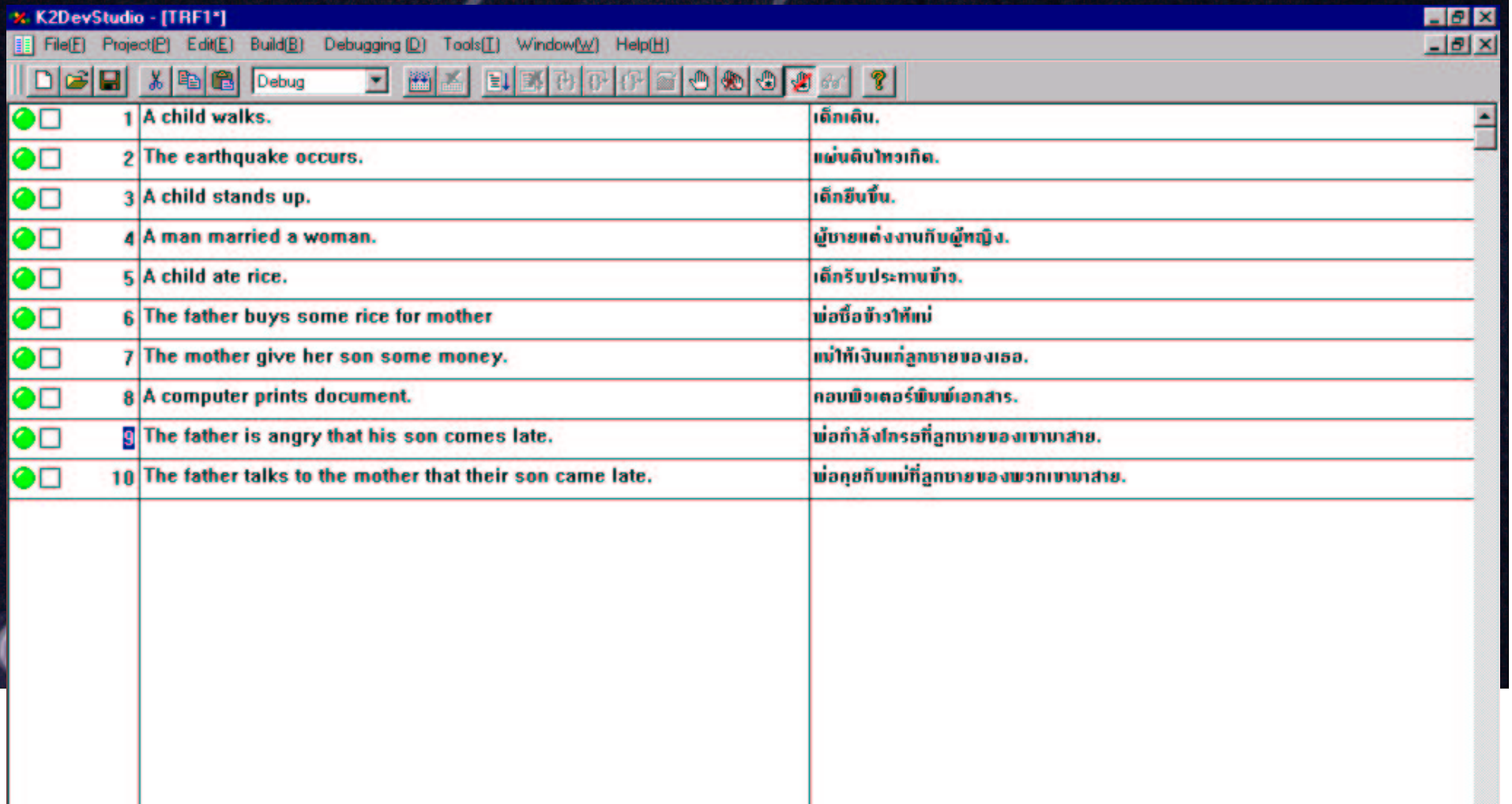
- Sentence Corpus for Evaluation
  - 770 Sentences
  - Designed by Japan Electronic Industry Development Association (JEIDA)
  - Have the characteristics for testing the following linguistic problems.
    - Word Level
      - Concept mismatching
      - Word absence
      - etc.
    - Sentence Level
      - Grammar
      - Modifier misplacement
      - etc.

Evaluation Result (number of sentence)

	Parsit	Smart Translator	AgentDict	PlaeThai	Translator 1.5
Perfect Translation Sentence	359	200	180	151	128
Partial Translation Sentence	351	504	486	465	567
False Translation Sentence	60	66	104	154	85
Total number of sentence	770	770	770	770	770
Accuracy	47-92%	26-91%	23-86%	20-80%	17-90%

ParSit 2 (Thai-English)

- Translation from Thai to English
- 13 Verb Patterns; 4,000 Rules; more than 30,000 words



1	A child walks.	เด็กเดิน.
2	The earthquake occurs.	แผ่นดินไหวเกิด.
3	A child stands up.	เด็กยืนขึ้น.
4	A man married a woman.	ผู้ชายแต่งงานกับผู้หญิง.
5	A child ate rice.	เด็กรับประทานข้าว.
6	The father buys some rice for mother	พ่อซื้อข้าวให้แม่
7	The mother give her son some money.	แม่ให้เงินแก่ลูกชายของเธอ.
8	A computer prints document.	คอมพิวเตอร์พิมพ์เอกสาร.
9	The father is angry that his son comes late.	พ่อกำลังโกรธที่ลูกชายของเขามาสาย.
10	The father talks to the mother that their son came late.	พ่อคุยกับแม่ที่ลูกชายของพวกเขามาสาย.

- Sansarn */san-sarn/*
  - สรรสาร (information search)
  - Language-independent probabilistic full-text search.
  - www.sansarn.com

Meaningful Bits

ADLTSUGKNOWLEDGEBWGZKTILA

ปร้จตเสีศฐาดึความรู้ษะภูกฮเศฉ

Mutual Information

$$Lm(xyz) = \frac{P(xyz)}{P(x)P(yz)}$$

$$Rm(xyz) = \frac{P(xyz)}{P(xy)P(z)}$$

x | **y** z

x **y** | z

where x is the leftmost character of string x**y**z
y is the middle substring of x**y**z
z is the rightmost character of string x**y**z
p( ) is the probability function.

High mutual information implies that x**y**z co-occurs more than expected by chance. If x**y**z is a word then its Lm and Rm must be high.

…Efunction… vs ...Function...

Entropy

$$LEnt(y) = -\sum P(xy/y) \cdot \log P(xy/y)$$

x **y**

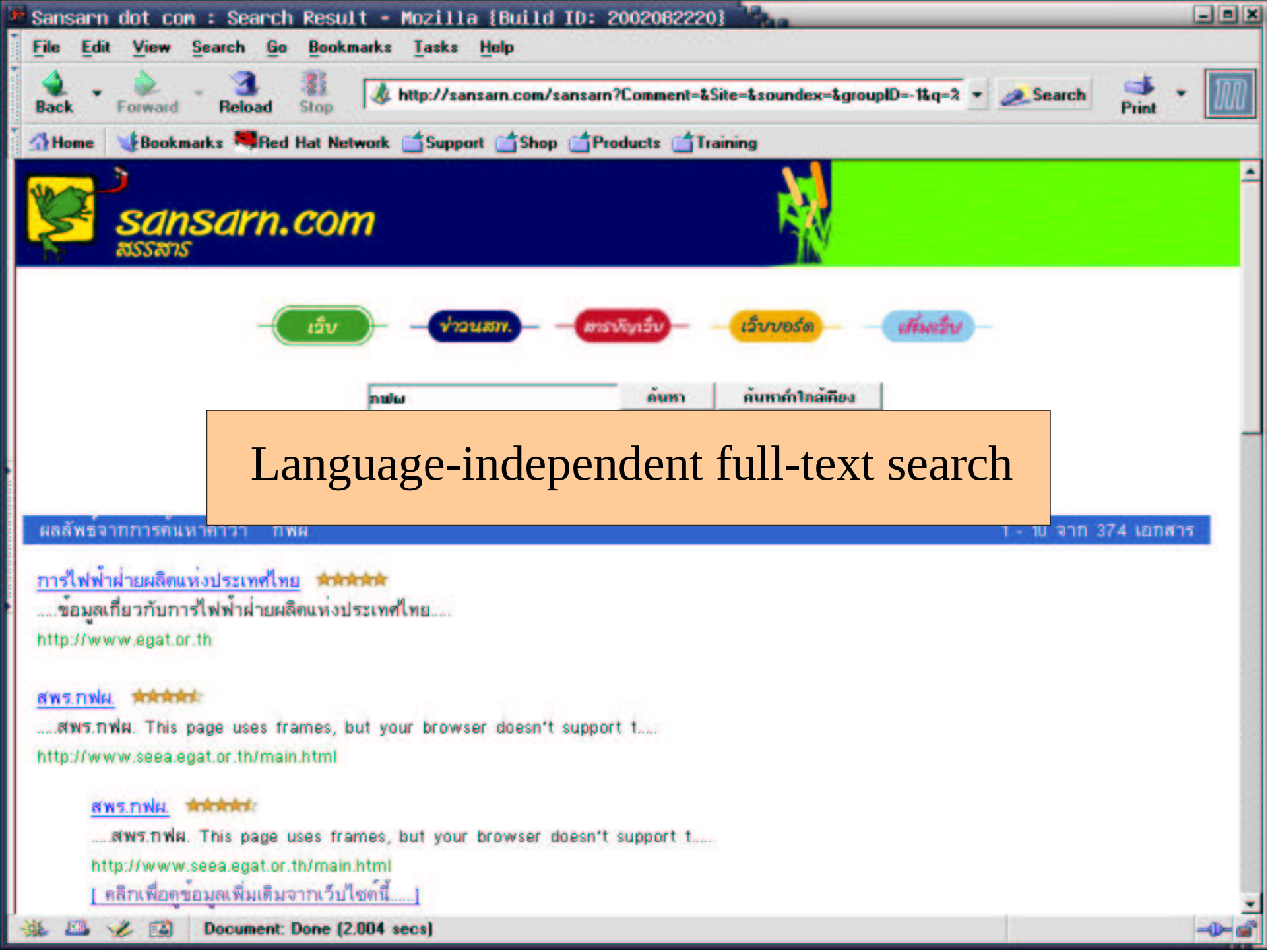
$$REnt(y) = -\sum P(yz/y) \cdot \log P(yz/y)$$

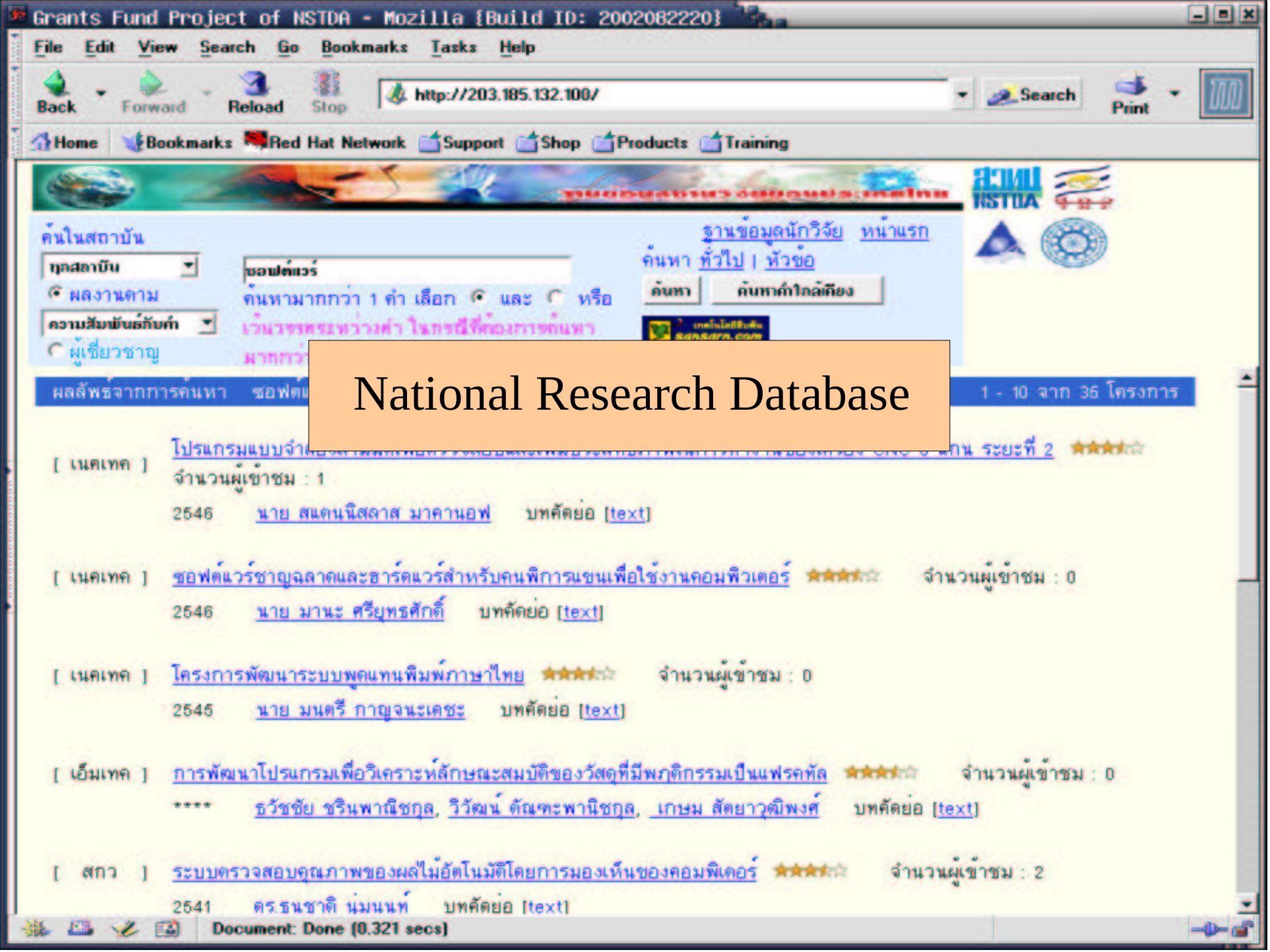
y z

where A is the set of characters
x is the leftmost character of string xyz
y is the middle substring of xyz
z is the rightmost character of string xyz
p( ) is the probability function.

Entropy shows the variety of characters before and after a word.
If y is a word then its left and right entropy must be high.

...?function... vs ...?unction...


Sansarn dot com : Search Result - Mozilla {Build ID: 2002082220}
File Edit View Search Go Bookmarks Tasks Help
Back Forward Reload Stop
http://sansarn.com/sansarn?Comment=&Site=&soundex=&groupID=-1&q=
Search Print
Home Bookmarks Red Hat Network Support Shop Products Training
sansarn.com
สรรสาร
เว็บ
ค้นหา
ค้นหาคำใกล้เคียง
Language-independent full-text search
การไฟฟ้าฝ่ายผลิตแห่งประเทศไทย
....ข้อมูลเกี่ยวกับการไฟฟ้าฝ่ายผลิตแห่งประเทศไทย....
http://www.egat.or.th
สพร.กฟผ.
....สพร.กฟผ. This page uses frames, but your browser doesn't support t....
http://www.seea.egat.or.th/main.html
สพร.กฟผ.
....สพร.กฟผ. This page uses frames, but your browser doesn't support t....
http://www.seea.egat.or.th/main.html
[คลิกเพื่อดูข้อมูลเพิ่มเติมจากเว็บไซต์นี้.....]
Document: Done (2.004 secs)


Grants Fund Project of NSTDA - Mozilla {Build ID: 2002082220}
File Edit View Search Go Bookmarks Tasks Help
Back Forward Reload Stop
http://203.185.132.100/
Search Print
Home Bookmarks Red Hat Network Support Shop Products Training
ค้นในสถาบัน
ทุกสถาบัน
ผลงานตาม
ความสัมพันธ์กับคำ
ผู้เชี่ยวชาญ
ซอฟต์แวร์
ค้นหามากกว่า 1 คำ เลือก และ หรือ
ฐานข้อมูลนักวิจัย หน้าแรก
ค้นหา ทั่วไป | หัวข้อ
ค้นหา
ค้นหาคำใกล้เคียง
National Research Database
ผลลัพธ์จากการค้นหา
1 - 10 จาก 35 โครงการ
[เนคเทค] ระยะที่ 2
จำนวนผู้เข้าชม : 1
2546 นาย สแตนนิสลาส มาคานอฟ บทคัดย่อ [text]
[เนคเทค] ซอฟต์แวร์ชาญฉลาดและฮาร์ดแวร์สำหรับคนพิการแขนเพื่อใช้งานคอมพิวเตอร์
จำนวนผู้เข้าชม : 0
2546 นาย มานะ ศรียุทธศักดิ์ บทคัดย่อ [text]
[เนคเทค] โครงการพัฒนาระบบพูดแทนพิมพ์ภาษาไทย
จำนวนผู้เข้าชม : 0
2545 นาย มนตรี กาญจนะเดชะ บทคัดย่อ [text]
[เอ็มเทค] การพัฒนาโปรแกรมเพื่อวิเคราะห์ลักษณะสมบัติของวัสดุที่มีพฤติกรรมเป็นแฟรคทัล
จำนวนผู้เข้าชม : 0
**** ธวัชชัย ชรินพาณิชกุล, วิวัฒน์ ตัณฑะพานิชกุล, เกษม สัตยาวุฒิพงศ์ บทคัดย่อ [text]
[สกว] ระบบตรวจสอบคุณภาพของผลไม้อัตโนมัติโดยการมองเห็นของคอมพิเตอร์
จำนวนผู้เข้าชม : 2
2541 ดร.ธนชาติ นิ่มนนท์ บทคัดย่อ [text]
Document: Done (0.321 secs)

ฐานข้อมูลนักวิจัย
หน้าแรก
รายชื่อทั้งหมด
รายชื่อตามสาขา

ล็อกอินเข้าระบบ
ชื่อ Login
รหัสผ่าน
เข้าระบบ
ลืมรหัสผ่าน | ลงทะเบียนใหม่

สาขางานวิจัย
Agricultural Science
Agricultural Produ
Agroindustrial com
Agronomy
Animal Breeding Ir
Animal Feed
Animal Husbandry
Animal Nutrition
Apiculture
Aquaculture

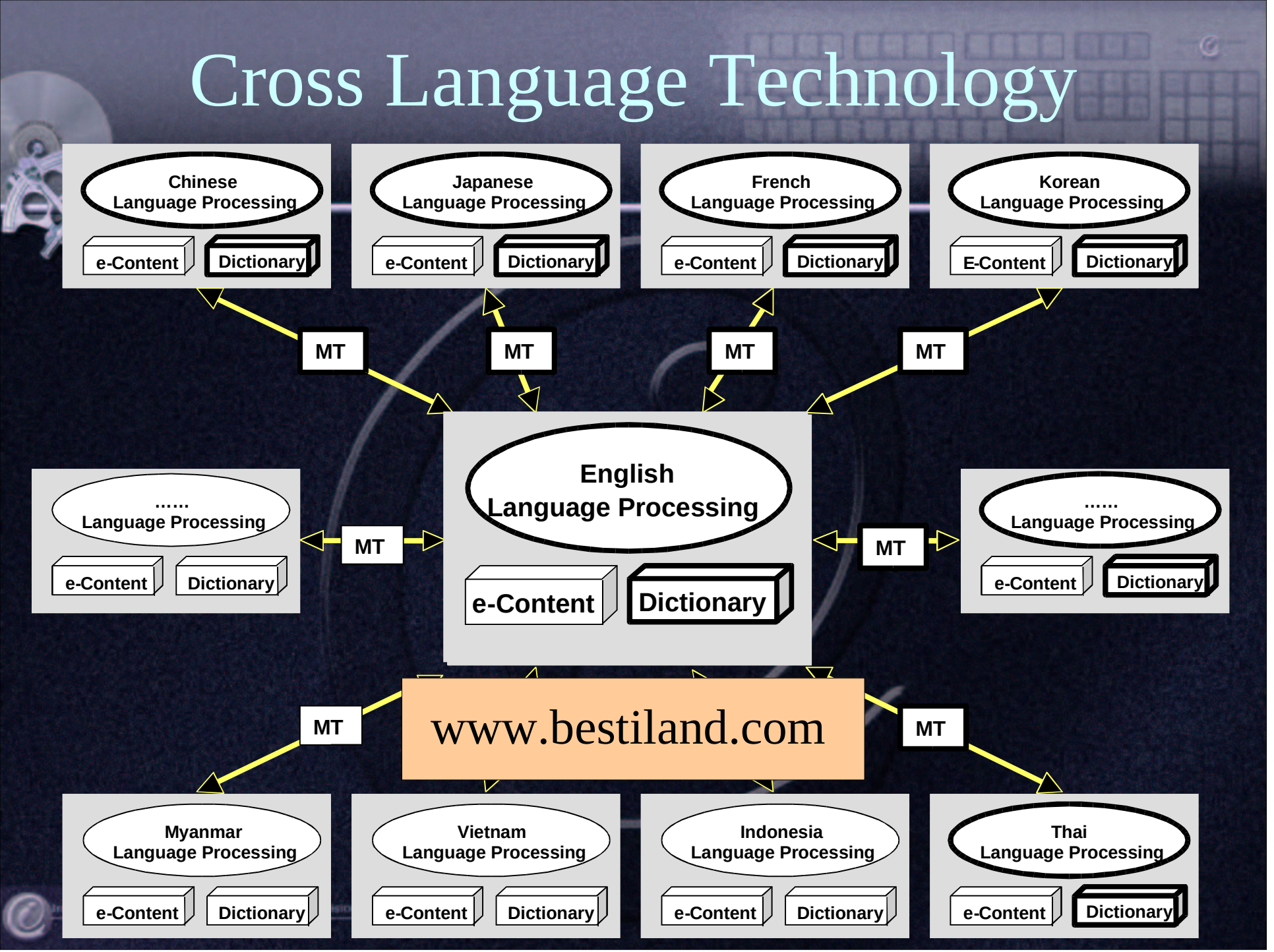
ค้นหา : ค้นหา ทั่วไป | หัวข้อ
ค้นหา

Researcher Database

รายชื่อนักวิจัยสาขา Agricultural Science click ที่ชื่อเพื่อดูรายละเอียด

ชื่อไทย	ชื่ออังกฤษ	สาขาวิจัยที่เชี่ยวชาญ
กัณยารัตน์ สุไพบูลย์วัฒน		PLANT BIOTECHNOLOGY/ HORTICULTURE/ AGRICULTURE
กิจการ ศุภมาตย์		P IMMUNITY INFECTIOUS MALS
จิระชัย กาญจนพฤติพงศ์	Kanjanaphumphong	NCE /ANIMAL NUTRITION/RUMINANT NUTRITION /RUMEN ECOSYSTEM /RUMEN EFFICIENCY
ชัยฤกษ์ มณีพงษ์		TISSUE CULTURE/AGRONOMY/PLANT BREEDING/PLANT GENETICS/GENETIC ENGINEERING/PROTOPLASTS
ประสาทพร สมิตะมาน	Prasartporn Smitamana	BOTANY/PLANT GENETICS TISSUE CULTURE/GENETIC ENGINEERING PROTOPLAST SHORTICULTURE TEMPERATE CUT FLOWER PLANT PROPAGATION PLANT IMPROVEMENT
ปรานอม พฤฒพงษ์		HORTICULTURE/TISSUE CULTURE
เปี่ยมศักดิ์ เมนะเศวต	Piamsak Menasveta	MARINE/AQUATIC/AQUACULTURE/AQUACULTURE BIOTECHMISTOGY/MARINE/OLLUTION/WATER RECIRCULATING SYSTEM/SHRIMP
พันทิพา พงษ์เพียจันทร์	Puntipa Pongpiachan	ANIMAL NUTRITION
ไพเราะ ทิพยทัศน์		NITROGEN FIXATION BIOCHEMISTRY/VITAMIN METABOLISM STEROID HORMONES CANCER REPRODUCTIVE SYSTEM RICENITROGEN FIXATION
ยุวดี มานะกาษม	Yuvadee Manakasem	HORTICULTURE/TISSUE CULTURE

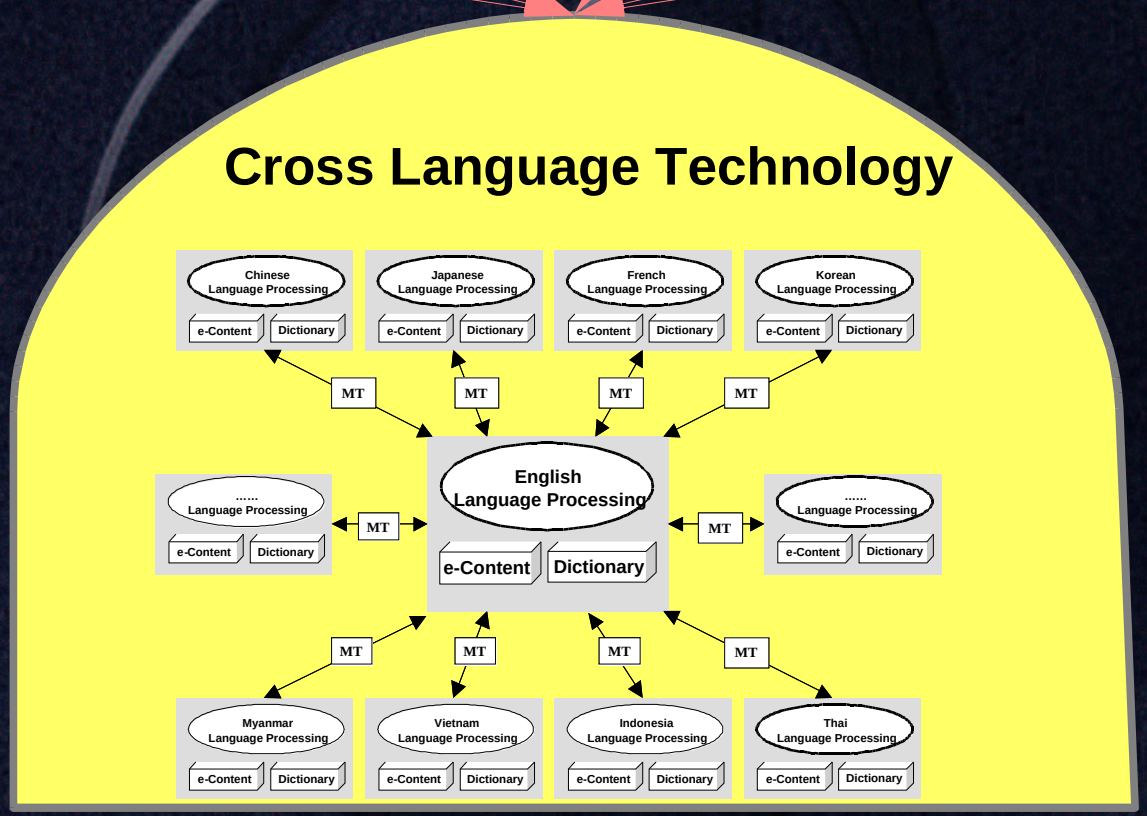
Cross Language Technology



Application over the Cross Language Tech.

E-services

Presentation | Extraction | Retrieval | Summarization | MT | Mining | Visualization



Cross Language Information Retrieval