KUI : The OSS-Styled Knowledge Development System

Virach Sornlertlamvanich
Thai Computational Linguistics Laboratory (TCL), NICT
virach@tcllab.org

The Association of Thailand Open Source Federation (TOSF.org)
To Manage the Knowledge

- Knowledge is managed by the knowledge users
- Knowledge is dynamically changed
- Knowledge is developed in an individual or a community
- Knowledge is both explicit and tacit
Types of Knowledge Development

- Topic of interest
  ... Initiation ...

- Opinion
  ... Discussion ...

- Localization
  ... Adoption ...

- Public-Hearing
  ... Criticism ...
Knowledge Development Process

- Topic of Interest
- Localization
- Opinion
- Public-Hearing

The 7th AOSS Symposium, Kuala Lumpur, Malaysia, 6-8 March 2006
Design of KUI

Knowledge User Interface for Knowledge Unifying Initiative

- GUI for knowledge engineering
- Online community-based discussion
- OSS-styled collaboration
- Multi-lingual solution
- Record of intention i.e. Chatlog
Features in KUI

- Opinion vs. Public-Hearing
  -> POLL

- Individual vs. Group works
  -> CHAT

- Record of Intention
  -> CHATLOG

- Selectional preference
  -> VOTE
KUI is a GUI for knowledge engineering. It provides a web interface accessible for pre-registered members. A contributor can comfortably move around in the virtual space from desk to desk to participate in a particular task. A working desk can be a meeting place for collaborative work that needs discussion through the ‘chat’, or allow a contributor to work individually by using the message slot to record each own comment. The working space can be expanded by closing the unnecessary frames so the contributor can concentrate on the task. All working topics can be statistically viewed through the provided tabs.

These tabs help contributors to understand KUI in the aspects of the current status of contribution and the tasks. A knowledge community can be formed and can efficiently create the domain knowledge through the features provided by KUI. These KUI features fulfill the process of human thought to record the knowledge.
The 7th AOSS Symposium, Kuala Lumpur, Malaysia, 6-8 March 2006

Opinion Poll

- Post your new opinion
- Vote for exist opinion
- Ranking opinion
- Realtime chatting

KUI's Topology

Topic of interest
The topic will be posted to draw the intention from the participants. The selected topics will then be further discussed in the appropriate step.

Opinion Poll
The selected topic is posted to call for opinions from the participants in this step. Opinion poll is conducted to get the population of each opinion. The result of the opinion poll provides the variety of opinions that reflects the current thought of the communities together with the consensus to the opinions.

Localization
Translation is the straightforward implementation of the localization. Collaborative translation helps producing the knowledge in multiple languages in the most efficient way.

Public-Hearing
The result of discussion will be revised and confirmed by gathering the opinions to the final draft of proposal.

Your Profile
Name: Virach Somlertlamvanich
Gender: Male
Country: Thailand
Reg. Date: [Change your profile]
Localization: Medical Questioning Translation

Where does it hurt?

Permissions

1101 รู้สึกเจ็บไหม
ความเห็น: ★★★★★
nartdanu: 13 January 2006 09:12:27

1104 哪里受伤了?
ความเห็น: ★★★★★
yue: 13 January 2006 09:12:27

<table>
<thead>
<tr>
<th>User</th>
<th>Chat Message</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>nartdanu</td>
<td>Can you see now. We have to click start button when we agree to move to next sentence</td>
<td>7 June 2005, 15:17:48</td>
</tr>
<tr>
<td>yue</td>
<td>oh, I can see</td>
<td>7 June 2005, 15:18:27</td>
</tr>
<tr>
<td>yue</td>
<td>that is great</td>
<td>7 June 2005, 15:18:38</td>
</tr>
<tr>
<td>nartdanu</td>
<td>Yeah :D</td>
<td>7 June 2005, 15:18:38</td>
</tr>
<tr>
<td>nartdanu</td>
<td>I finished my translation and ready to move.</td>
<td>7 June 2005, 15:19:05</td>
</tr>
<tr>
<td>yue</td>
<td>ok, go ahead</td>
<td>7 June 2005, 15:19:56</td>
</tr>
<tr>
<td>nartdanu</td>
<td>Ok</td>
<td>7 June 2005, 15:20:03</td>
</tr>
<tr>
<td>yue</td>
<td>it's a little slow</td>
<td>7 June 2005, 15:20:42</td>
</tr>
<tr>
<td>nartdanu</td>
<td>I see because now i cannot press the start button, wait for a while</td>
<td>7 June 2005, 15:21:16</td>
</tr>
<tr>
<td>yue</td>
<td>ok</td>
<td>7 June 2005, 15:21:30</td>
</tr>
<tr>
<td>nartdanu</td>
<td>Please select confidence, then we can press start button</td>
<td>7 June 2005, 15:23:41</td>
</tr>
<tr>
<td>yue</td>
<td>ok</td>
<td>7 June 2005, 15:24:00</td>
</tr>
<tr>
<td>kergrit</td>
<td>หนึ่งถึงสี่</td>
<td>9 January 2006, 14:06:55</td>
</tr>
</tbody>
</table>
### Localization: Medical Questioning Translation

#### 11) When you cough

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Language</strong></td>
<td><strong>Translation</strong></td>
</tr>
<tr>
<td>Thai</td>
<td>เมื่อคุณไอ</td>
</tr>
<tr>
<td>Chinese</td>
<td>你什么时候咳嗽？</td>
</tr>
</tbody>
</table>

#### 12) Do you cough?

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Language</strong></td>
<td><strong>Translation</strong></td>
</tr>
<tr>
<td>Thai</td>
<td>คุณไอไหม</td>
</tr>
<tr>
<td>Thai</td>
<td>คุณไอปาน</td>
</tr>
<tr>
<td>Chinese</td>
<td>你咳嗽吗？</td>
</tr>
</tbody>
</table>
Localization: Computer Dictionary Translation

คำอธิบาย: แปลคำศัพท์คอมพิวเตอร์

คำค้น: access

ภาษา: All

ค้นหาตามตัวอักษร: ABCDEFGHIJKLMNOPQRSTUVWXYZ

รายการของความคิดเห็น

1) access code
   ภาษา: Thai
   ความคิดเห็น: รหัสการเข้าถึง

2) access method
   ภาษา: Malay
   ความคิดเห็น: kaedah capaian
   ภาษา: Thai
   ความคิดเห็น: วิธีเข้าถึง

3) access time
   ภาษา: ความคิดเห็น:
Localization: NICT's Japanese-English News Articles Alignment

The Japanese and English translated sentences are selected from the news articles of two Japanese newspapers, i.e., Yomiuri Shimbun and Daily Yomiuri, ranging from September 1989 to December 2001. The total number of the articles is 59,086. The Japanese and English translation is the result from automatic alignment. For the details, see Masao Utsumi and Hitoshi Isahara. (2003) Reliable Measures for Aligning Japanese-English News Articles and Sentences, ACL-2003, pp. 72-79. The original Japanese-English aligned sentences can be downloaded at http://www2.nict.go.jp/jt/a132/members/mutsumi/jea/index.html. There are totally 150,000 sentences of the result of one-to-one sentence alignment (JENAAD-p11).

Number: 100 items

<table>
<thead>
<tr>
<th>Number</th>
<th>Japanese</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>1)</td>
<td>欧州は、エディンバラにおいて合意され、コペンハーゲンにおいて強化された成長イニシアチブを精力的に実行しつつある。</td>
<td>Europe is carrying out vigorously the Growth Initiative agreed in Edinburgh and strengthened in Copenhagen.</td>
</tr>
<tr>
<td>2)</td>
<td>我々は、ロシアの経済発展にとって、改善された市場アクセスが重要であることを認識する。</td>
<td>We recognize the importance of improved market access for economic progress in Russia.</td>
</tr>
<tr>
<td>3)</td>
<td>法人レベルでのパートナーシップ及びマネージメント支援は、特に効果的であるのである。</td>
<td>Restructuring and management assistance, at corporate level can be particularly effective.</td>
</tr>
</tbody>
</table>
KUI is not ...

• SourceForge.net  
  ... for its knowledge/document development support

• Wiki.org  
  ... for its record of intention and selectional preference

• Weblog  
  ... for its well-structured, domain-oriented design
Evaluation and On-going Tasks

Localization

• Medical question translation Experiment in ICE2005 among China, France, Korea, Japan, Malaysia, Thailand.

• Technical term dictionary translation (computer)

• NICT's Japanese-English news article
Call for Contribution

KUI is an Open Source styled collaborative knowledge development.

KUI
www.tcllab.org/kui
Potential Partnerships

- Language Observatory
- African Academy of Languages (ACALAN)
- Linguasphere Observatory
- UNESCO's World Network for Linguistics Diversity
- University of Melbourne
KUI-based Applications

- Thai Computational Lexicon (TCLLEX)
- Multi-lingual Search Engine
- Language Identifier on Byte Stream (LIBS)
- Web Language Engineering (WLE)
ResultOfAction: a kind of abstract thing referring to that is a product of human activity.
<table>
<thead>
<tr>
<th>Word ID</th>
<th>Thai Word</th>
<th>English Word</th>
<th>Lexicographer</th>
<th>Date</th>
<th>Status</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>นน</td>
<td>cuddle</td>
<td>pop</td>
<td>01 July 2005 09:21:55</td>
<td>Assigned</td>
<td>view, edit, Delete</td>
</tr>
<tr>
<td>2</td>
<td>นน</td>
<td>cuddle</td>
<td>pop</td>
<td>01 July 2005 09:22:46</td>
<td>Assigned</td>
<td>view, edit, Delete</td>
</tr>
<tr>
<td>3</td>
<td>นน</td>
<td>reed</td>
<td>pop</td>
<td>01 July 2005 09:23:30</td>
<td>Assigned</td>
<td>view, edit, Delete</td>
</tr>
<tr>
<td>4</td>
<td>นน</td>
<td>hatch</td>
<td>pop</td>
<td>01 July 2005 09:23:30</td>
<td>Assigned</td>
<td>view, edit, Delete</td>
</tr>
<tr>
<td>5</td>
<td>กลางอก</td>
<td>embrace</td>
<td></td>
<td>01 July 2005 09:23:30</td>
<td>Approved</td>
<td>view, edit, Delete</td>
</tr>
<tr>
<td>6</td>
<td>กดหู</td>
<td>ear</td>
<td></td>
<td>01 July 2005 09:23:30</td>
<td>Submitted</td>
<td>view, edit, Delete</td>
</tr>
<tr>
<td>7</td>
<td>กะเสา</td>
<td>base</td>
<td></td>
<td>01 July 2005 09:23:30</td>
<td>Assigned</td>
<td>view, edit, Delete</td>
</tr>
<tr>
<td>8</td>
<td>กะเสือ</td>
<td>reed</td>
<td></td>
<td>01 July 2005 09:23:30</td>
<td>Assigned</td>
<td>view, edit, Delete</td>
</tr>
<tr>
<td>9</td>
<td>กะไข่</td>
<td>incubate</td>
<td>pop</td>
<td>01 July 2005 09:23:30</td>
<td>Submitted</td>
<td>view, edit, Delete</td>
</tr>
<tr>
<td>10</td>
<td>กะไม้</td>
<td>trunk</td>
<td></td>
<td>01 July 2005 09:23:30</td>
<td>Assigned</td>
<td>view, edit, Delete</td>
</tr>
<tr>
<td>11</td>
<td>กษัตริย์</td>
<td>regalia</td>
<td></td>
<td>01 July 2005 09:23:30</td>
<td>Problem Submitted</td>
<td>view, edit, Delete</td>
</tr>
</tbody>
</table>
Multilingual Search Engine (Experimental)

Select Language  Japanese  北京  Search  Soundex Search

Search Result:

ディレクトリ別一覧 ('R'で始まるホームページ)  12837 tf=3
....  ruoyue—中央民族大学非公式通信 中国、北京にある中央民族大学のホームページです。留学情報のほか、北京の最新の話題や中国旅行情報なども扱っていますので、留学する方、している方、もうすぐ北京に行かれる方は覗いてみてください。  ryuboh838—ryuboh838・ryupara—りゅうばらだいす ぼくの名前はリュウ。犬・雑種・♂・6歳。ぼくが日頃思っていることをつづって...
http://www1.plala.or.jp/town/PH_menu_index.R.htm

リンククラブニュースダイジェスト  11359 tf=1
.... 空港と上海を結ぶリニアモーターカーは時速450Kmで走行。現在は車で約1時間ちょっとかかる距離だがそうですが、リニアモーターカーだと約8分で上海へ到着するそうです。この工事が成功すると、次は北京まで路線を延ばすとのこと。中国では、基本的に土地は国家のもの。その気になれば物凄い速度で大規模開発が可能なんですね。中国恐るべき...
http://info.linkclub.or.jp/digest/2003_01/09.html

映画評  11803 tf=1
....「女」としての愛人を囲う、との。うーん。映画の主旨をけんしきれない発言である。 北京ヴァイオリン ‘02 中
03/06/15 浜谷さ・シネマ バイオリンの得意な少年チュンは父リュに連れられて北京へ上京。リュは息子の才能を伸ばそうとさき講
Multilingual Search Engine (Experimental)

Select Language  Chinese  北京  

Search Result:  北京

1 - 10 from 2278 documents (3.39 sec)

人民网-北京视窗  7227  tf=25
...园林静思园 &gt;更多 区县网站 东城区 西城区 崇文区 宣武区 朝阳区 海淀区 丰台区 石景山 门头沟 昌平区 通州区 顺义区 延庆县 房山区 大兴区 怀柔区 密云区 平谷区 人民地方网 北京 天津 山西 河北 内蒙古 湖北 湖南 河南 台湾 香港 澳门 辽宁 吉林 黑龙江 广东 广西 海南 新疆 青海 宁夏 陕西 甘肃 上海 安徽 浙江 福建 江苏 江西 西藏 四川 云南 贵州....
http://bj.people.com.cn/

中国旅游网  4552  tf=19
...个环节上来，使货畅其流，使旅游消费者的需要，得到充分的满足。为此，必须克服商品销售上的"两多两少"：旅游商品市场中一般性的，全国哪里都能买到的商品多；一些景点景区销售粗制滥造的、无北京特色、无景点特色的旅游商品多。在市区、地区级商业中心，旅游商品的专营店和专营柜台少；在景点景区中的商店和摊点，销售与景点特色有关联的旅游商品少。为了使旅游商品货畅其流，必须建立和健全旅游商品....
http://www.cnta.gov.cn/32-lydy/2j/12-1.asp

&amp;#27472;&amp;#27954;&amp;#22291;&amp;#26126;&amp;#32593;  8291  tf=15
...技侣对付和平的请愿者，以暴力逮捕学 員並接著透過媒體醜化學員。 二百遊客目睹愛爾蘭西人弟子天安門正義之舉（圖） [19.03.2002] 吉瑞德.
Multilingual Search Engine (Experimental)

Search Result: samsung

1 - 10 from 22 documents (1.13 sec)

KNTO - PCO 13548 tf=4
Fax 02-6000-1306 Conference Interpretation & Translation Institute In Asia Pacific - Rm.1110, Samsung Jeil Bldg., 702-2, Yeoksam-dong, Gangnam-gu, Seoul 135-080 Tel 02-527-2940 Fax 02-527-2942 ....
http://www.knto.or.kr/eng/06_convention/06_04.html

드림위즈 취업 17123 tf=3
....견 상세검색 조회수:1202 (2004/12/7 14:01 기준) (주)휴먼코아 : 삼성물산 SAMSUNG MALL 웹디자이너 / 사무사원 모집 (주5일) (주)휴먼코아 삼성물산 SAMSUNG MALL .. · My scout에 스크랩하기 · 친구에게 이-메일 보내기 · 관심기업으로 등록하기 · 프린트 하기 · 스크랩한 회원보기 · 지원현황통계 기업분석 ....
http://dreamwiz.scout.co.kr/jobs/all/contentview.asp?id=359413

드림위즈 취업 17775 tf=2
....도소재 대기업 웹디자이너 모집 매칭효스탁(주) 1~3년 무관 12/11(토) 12/04 [급]Web Design경력사원모집 사람과기술 1~3년 무관 채용시까지 12/04 삼성물산 SAMSUNG MALL 웹디자이너 / 사무사원 모집 (주5.. (주)휴먼코아 무관 무관 채용시까지 12/04 삼성물산 SAMSUNG MALL 웹디자이너 / 사무사원 모집 (주5......
Multilingual Search Engine (Experimental)

Select Language:  

Search Result:  กรุงเทพ 1 - 10 from 316 documents (1.26 sec)

Nation Qvote 1293 tf=2

.... 09:13 กระทบเรือนท่านชาย บุหรี่ของท่านน่วนมีทางการเมืองเลยครับ ท่านข้าราชการมากแล้ว คณบดีประกาศยกเว้นกรุงเทพ อยากให้ท่านเป็นต้องแต่ เนื้อข้าคิดว่ากรุงเทพคือประเทศไทยและข้าไม่ชอบท่าน เลิกเละครับ ให้คุณ....

Nation Qvote 1302 tf=2

....เมื่อคืนนี้dek love phelps 30/08/04 21:38 น.ไม่ได้คุยกันไม่ได้แล้วไปซื้อ ซื้อของ สำนัก ทำให้คุณซื้อนานกรุงเทพตั้ง4ปี ตั้งแต่ควันเสือกตั้ง ถามไม่ เลิกถาม的人 แม้ว่าก็ไม่ทำอะไรก็ไม่ได้ยุ่งแล้ว วันนี้...   

กายใจ - Body-Heart (The Krungthep turakij web site) 1422 tf=2

.... อะไรที่ในประเทศดูยังคง สอบถามโทร 0-9032-2921 หรือ 0-1751-4456 สอบถาม...คณะบริหารธุรกิจ มหาวิทยาลัยกรุงเทพ จัดสัมมนาเรื่อง "quot; การบริหารด้านทุน - ทางของสี่ส่วน SMEs"quot; ระหว่างวันที่ 16 ธันวาคม 254....

www.thairath.co.th 37 tf=1

Done
LIBS: Language Identifier on Byte String

Result
Hindi

Input string
एशिया में बह फ्लू के

Text samples including 60 languages
Afrikaans
die heilige graal van die

Albanian
provat në të gjithë botën
Motivation:

- Establishing Thailand-Japan collaborative archiving on broadband Internet (NECTEC-MIC 45M bps high speed link)

- Collaboration between TCL-Thailand and NUT-Japan to support Language Observatory Project and Asian Language Resource Network

- Real-time monitoring population of web document
Web Language Engineering
for Open Collaborative Archiving (WLE-OCA)

Purpose:

- Providing realtime monitoring system of the collaborative crawling

- Providing statistics of the existing of web documents in the aspects of language, encoding and the defined charset to raise awareness about the minority languages in the online community, and to encourage the development of web documents in the less-computerized languages

- Archiving the web documents for web language study to prevent the fall-off in using of the minority languages in the online community
Goal:

- Archiving web document at 1M pages/day/site
- Enhancing Language ID from the current 86 to 200 language-encoding pairs
- Evaluating Multi-lingual and Cross-lingual search engine of the archive
Web Language Engineering for Open Collaborative Archiving (WLE-OCA)

About WLE-OCA

It is said that there are about 7,000 languages using all over the world. However, not all of the languages are used in the online community. About 35% of the online language populations are using English, and about 68% of web contents are written in English as reported by the GLocal Reach in September 2004. Many efforts have been making to prevent the fall-off in using of the minority languages in the online community and the less computerized languages. For the minority online languages, it needs a set of useful software utilities to facilitate the native users. The efficiency in web document development and utilization such as authoring and retrieving tools, is basically required to sustain the population. For the less computerized languages, standardization of the encoding system is crucial. It is the first platform for the languages to participate in the online community.

Web Language Engineering (WLE) is established to study about the techniques in processing the languages using in publishing web documents. We need a timely report of the web language population to monitor results of collaborative works among the distributed Nodes of Service (NOS). The collaborative works are realized to balance the load based on the time proximity between the target server and NOS. The shortest access time determines the NOS in charge of the target server. In each NOS, language identification and web document archiving consequently work on the web documents for efficient retrieval.

The techniques in WLE supports an Open Collaborative Archiving (OCA) in the following aspects:

- Collaborative crawling
- Language identification
- Multi-lingual and cross-lingual search engine

Motivation
Architecture:

Time proximity collaborative crawling

Japanese

Thai

Research Network

Public Network

Language Identification

Database

Indexer

Archive Search

Monitoring

(MIC-NECTEC 45 Mbps)
WLE Component:

- Time proximity collaborative crawler
- Language identification on byte string
- Multilingual search engine
Web Language Engineering
for Open Collaborative Archiving (WLE-OCA)

Accumulated web documents as of 2005-11-28

<table>
<thead>
<tr>
<th>Name</th>
<th>Start crawl date</th>
<th>Total page</th>
<th>Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>TH-ORCHID @ TCL</td>
<td>2005-11-18 (11 day ago)</td>
<td>3,376 pages</td>
<td>307 pages per day</td>
</tr>
<tr>
<td>JP-KIKU @ NUT</td>
<td>2005-11-21 (8 day ago)</td>
<td>7,222 pages</td>
<td>903 pages per day</td>
</tr>
<tr>
<td>JP-SAKURA @ NUT</td>
<td>2005-08-05 (116 day ago)</td>
<td>337,242 pages</td>
<td>2,907 pages per day</td>
</tr>
</tbody>
</table>

*We limit the crawling to ccTLD of Asia region, excluding CJK(.cn, .jp, .kr).*
### Distribution of Encoding for Language:

<table>
<thead>
<tr>
<th>Encoding</th>
<th>Pages</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISO-8859-1</td>
<td>201,752</td>
<td>100.00</td>
</tr>
<tr>
<td>total</td>
<td>201,752</td>
<td>100.00</td>
</tr>
</tbody>
</table>

### Distribution of Charset for Language:

<table>
<thead>
<tr>
<th>Charset</th>
<th>Pages</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISO-8859-1</td>
<td>135,912</td>
<td>67.37</td>
</tr>
<tr>
<td>Windows-1252</td>
<td>19,585</td>
<td>9.71</td>
</tr>
<tr>
<td>utf-8</td>
<td>8,902</td>
<td>4.41</td>
</tr>
<tr>
<td>ISO-8859-9</td>
<td>7,079</td>
<td>3.51</td>
</tr>
<tr>
<td>Windows-1254</td>
<td>6,083</td>
<td>3.02</td>
</tr>
<tr>
<td>windows-1255</td>
<td>6,059</td>
<td>3.00</td>
</tr>
<tr>
<td>windows-874</td>
<td>5,642</td>
<td>2.80</td>
</tr>
<tr>
<td>iso-8859-8</td>
<td>2,030</td>
<td>1.01</td>
</tr>
<tr>
<td>NA</td>
<td>1,979</td>
<td>0.98</td>
</tr>
<tr>
<td>windows-1256</td>
<td>1,761</td>
<td>0.87</td>
</tr>
<tr>
<td>TIS-620</td>
<td>1,727</td>
<td>0.86</td>
</tr>
<tr>
<td>iso-8859-1</td>
<td>1,166</td>
<td>0.58</td>
</tr>
<tr>
<td>visual</td>
<td>1,132</td>
<td>0.56</td>
</tr>
<tr>
<td>windows-1251</td>
<td>619</td>
<td>0.31</td>
</tr>
<tr>
<td>tis-620</td>
<td>499</td>
<td>0.25</td>
</tr>
<tr>
<td>windows-1253</td>
<td>351</td>
<td>0.17</td>
</tr>
<tr>
<td>us-ascii</td>
<td>249</td>
<td>0.12</td>
</tr>
<tr>
<td>unicode</td>
<td>191</td>
<td>0.09</td>
</tr>
<tr>
<td>Shift_JIS</td>
<td>121</td>
<td>0.06</td>
</tr>
<tr>
<td>x-user-defined</td>
<td>90</td>
<td>0.04</td>
</tr>
</tbody>
</table>
Collaborative Crawler

Server List

Host: TH-ORCHID @ TCL
Start crawl date: 2005-11-18

Host: JP-KIKU @ NUT
Start crawl date: 2005-11-21

Crawler Status
Language Distribution
of 347,875 pages from all sites

<table>
<thead>
<tr>
<th>Language</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>37.3</td>
</tr>
<tr>
<td>Hebrew</td>
<td>30.4</td>
</tr>
<tr>
<td>Thai</td>
<td>15.3</td>
</tr>
<tr>
<td>Turkish</td>
<td>6.1</td>
</tr>
<tr>
<td>Norwegian</td>
<td>5.1</td>
</tr>
<tr>
<td>Italian</td>
<td>2.2</td>
</tr>
<tr>
<td>Russian</td>
<td>2.1</td>
</tr>
<tr>
<td>Malay</td>
<td>1.4</td>
</tr>
</tbody>
</table>

Log out
WLE is to learn about the origins, languages and contents of web documents aiming at preserving and populating the diversity of languages.

WLE-OCA

www.tcllab.org/weblang
Established in November 2002, Thai Computational Linguistics Laboratory (TCL) is a partnership-laboratory of Computational Linguistics Group (Japan), under Keihanna Human Info-Communication Research Center (KICR) of National Institute of Information and Communications Technology (NICT).

TCL is located at Thailand Science Park, in the building of National Electronics and Computer Technology Center (NECTEC), where many research and development on info-communications, materials, and bio-technologies are being conducted. It is commonly known that many Asian languages have their own unique problems. Existing methods in Natural Language Processing (NLP) may not be practical when they are directly applied to Asian languages. Customizing existing methods and developing new technologies have been TCL's main research activities over the past three years.