PanLex
Panlingual Lexical Collaboration

Jonathan Pool
University of Washington Computational Linguistics Laboratory
22 April 2008
Pre-Demo

PanLex (http://panlex.org/cgi-bin/panlex13.cgi)

Task 1.

You encounter the word “प्रधानाध्यापक”.

What language is it in?

What does it mean?

Task 2.

You encounter the word list at http://www.geonames.de/peace.html.

Is its content already in PanLex?

If not, how can you contribute it?
Goals

Facilitate panlingual:

4. Vigor.
3. Discursive intertranslatability.
2. Lexical intertranslatability.
1. Lexical collaboration.
Strategy

Goal 1: Facilitate panlingual lexical collaboration

How?

1. Assemble valuable panlingual data.
2. Make the data accessible.
3. Invite contributions to the data.
4. Localize the interface panlingually.
Tactics

Tactic 1: Assemble valuable panlingual data.

How?

Borrow data from TransGraph.

Expression (lexeme) equivalences from 357 dictionaries.

13 multilingual, 344 bilingual.

1050 languages.

2.5 million expressions.

8 million expression tokens.

Accept (mainly) TransGraph’s lightweight schema.

An expression is just a string in a language.

A meaning is just a source-specific ID.

A denotation is just a source assigning a meaning to an expression.

A translation is just 2+ denotations with the same meaning.
Example
TransGraph Data
Tactics

Tactic 2: Make the data accessible.

How?

Open-source (PostgreSQL) database (vs. TransGraph).
Perl CGI-DBI application to query and modify the data.
Domain "panlex.org" to access the application.
All data exposed (vs. PanImages).
Data retrievable interactively and by plain-text or XML file export.
Tactics

Tactic 3: Invite contributions to the data.

How?

User contributions nondestructive.
Not a Wiki, not moderated.

Contributable data:

[Language varieties (vs. TransGraph languages).]
Expressions.
Sources.
Denotations.

Contribution modes:
Batch (file upload; plain-text or XML).
Incremental (interactive editing).
Tactics

Tactic 4: Localize the interface panlingually.

How?

In vivo localization.

Interface entirely lemmatic.

Therefore, PanLex can translate the interface.

Translation core: developer-attested translations.

Translation periphery: election with sources voting.
Evaluation

Test 1 (expert user):
- 15 query and modification tasks with test questions.
  Failures and comments inspired interface changes.

Test 2 (expert user):
- Found, formatted, checked, and uploaded data from:
  - Nepali-Esperanto dictionary.
  - English-Yiddish dictionary.
  - Eight-language medical glossary.
Future Work

Coverage

Add dictionaries.
Recruit user-added dictionaries.
Add source types:
  Thesauri.
  WordNets.
  Library subject headings.
  Locale repositories.
  Monolingual resources.
Export additions to TransGraph.
<table>
<thead>
<tr>
<th>Language</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>428,550</td>
</tr>
<tr>
<td>Hungarian</td>
<td>264,927</td>
</tr>
<tr>
<td>French</td>
<td>172,435</td>
</tr>
<tr>
<td>German</td>
<td>135,505</td>
</tr>
<tr>
<td>Czech</td>
<td>110,623</td>
</tr>
<tr>
<td>Croatian</td>
<td>92,146</td>
</tr>
<tr>
<td>Spanish</td>
<td>82,503</td>
</tr>
<tr>
<td>Estonian</td>
<td>73,628</td>
</tr>
<tr>
<td>Italian</td>
<td>72,861</td>
</tr>
<tr>
<td>Esperanto</td>
<td>62,928</td>
</tr>
<tr>
<td>Arabic</td>
<td>56,122</td>
</tr>
<tr>
<td>Finnish</td>
<td>54,439</td>
</tr>
<tr>
<td>Japanese</td>
<td>50,247</td>
</tr>
<tr>
<td>Dutch</td>
<td>46,921</td>
</tr>
<tr>
<td>Portuguese</td>
<td>38,237</td>
</tr>
<tr>
<td>Russian</td>
<td>36,779</td>
</tr>
<tr>
<td>Breton</td>
<td>29,436</td>
</tr>
<tr>
<td>Serbian</td>
<td>27,072</td>
</tr>
<tr>
<td>Romanian</td>
<td>26,516</td>
</tr>
<tr>
<td>Kurdish</td>
<td>25,315</td>
</tr>
<tr>
<td>Swedish</td>
<td>24,495</td>
</tr>
</tbody>
</table>
Future Work

Features

More query functions.
User SQL entry.

Usability

Test and improve interface.
Non-expert interface.

Standards

Lemmatic forms (e.g., English “to”).
Multiword lexemes.