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

tur
dilbilim

ell
γλωσσολογία

eng
linguistics

est
keeleteadus
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.
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.