Back to the roots: building simple bilingual dictionaries

Mathieu MANGEOT
Condillac - LISTIC - Université de Savoie
F-73376 Le Bourget du Lac Cedex France
Mathieu.Mangeot@univ-savoie.fr

Overview

- Current Situation
- Identified Problems
- The proposal : building bilingual dicts
- Conclusion

Current Situation

- The Meetings
- The Data
- The Platform

The meetings

- 2000: Tokyo
- 2001: Grenoble
- 2002: Tokyo
- 2003: Sapporo
- 2004: Genève
- 2005: Chiang Rai

The Data in original structure

- 15 dictionaries
  - Monolingual: ThaiDict, DiCo
  - Bilingual: Armement, Cedict, Ding, EngKor, Fjocean, GDEF, JMDict, Kanjidic, Maniette, VietDict, WaDokuJiTen
  - Multilingual: FeM, ELRA database
- 9 languages
  - French, English, German,
  - Japanese, Chinese, Korean,
  - Malay, Thai, Vietnamese
- More than 1 million of entries

The data in Papillon structure

- 613 French lexies from DiCo
  - Converted 1 by me and 2 by Guy Lapalme
- 202 French, 183 English, 105 Japanese
  - Written from scratch by Mutsuko Tomokiyo
- 67 Malaysian lexies
  - Written from scratch by T. Enya Kong team
- 42 axies
  - Written from scratch by myself

The platform: history

- 2001: first prototype by Gilles Sérasset
- 2002-2003: development by myself + GS
- 2004: development of the editor + debug
- 2005: stable version used in 4 projects:
  - Papillon
  - GDEF: bilingual Estonian-French dict.
  - LexAlp: terminology deu, eng, fra, ita, slo
  - New project from a university in Sri-Lanka

The platform: annex tools

- Awstats: connection log analyser
  - perl script
- phpBB2: online forum
  - php package
- Uplug: aligned bilingual corpora
  - Mixture of perl, shell and tools

Identified Problems

1. Wrong bottleneck
2. Few Japanese collaborators
3. Macrostructure too abstract
4. Wrong entry unit
5. Microstructure too complex
6. Eyes bigger than stomach

1. Wrong bottleneck

- First though
  - An adequate software will solve it
- What happens
  - the Jibiki platform is fully operational
    - used by 3 projects
- My take
  - Lexicographical aspects underestimated

2. Few Japanese collaborators

- First though
  - Papillon = French-Japanese collaboration
- What happens
  - Only 2 Japanese collaborate
    - Mutsuko Tomokiyo & Kyoko Kuroda
- My take
  - Japanese don't need another F-J dict
  - We did not know Japanese lexicographers

3. Macrostructure too abstract

- First though
  - Build a dictionary with a pivot structure
- What happens
  - The structure is too difficult to understand
  - No example set available
- My take
  - Simplify the structure
  - Recuperate the PARAX data

4. Wrong entry unit

- First though
  - Use the lexie as the entry unit
- What happens
  - Data shared between lexies is not updated
  - No distinction between lexies of the same vocable and homograph vocables
  - Most people are used to the “vocable” entry unit
- My (+CB) take
  - Use the “vocable” entry unit
  - Create and store the tree hierarchy between lexies

5. Microstructure too complex

- First though
  - Use the DiCo microstructure (from MTT)
    - Very complex for non specialists
- What happens
  - Takes too much time to learn
- My take
  - People do not contribute if they do not understand the whole structure

6. Eyes bigger than stomach

- First though
  - We will build a skeleton dict by reusing existing data
- What happens
  - Many theoretical achievements
    - But no core data set available yet
- My take
  - Start now to build data “by hands”
  - => useful for bootstrapping and theory validation

The proposal: building simple bilingual dicts

- Structures
  - Macrostructure, entry unit, microstructure
- Methodology
  - Linking process
  - Writing protocol
- Reusable data
  - Monolingual
  - Bilingual

Macrostructure

- Keep a pivot macrostructure
  - An axie links 2 lexies of 2 languages
- But
  - The pivot structure is hidden
  - Contributors see only bilingual links

Entry unit

- “Vocable” entry unit
- One grammar block per part-of-speech
- Each block contains one or more lexies
- The lexies are sorted by frequency

Microstructure

- Tend to DiCo but more simple:
- For every vocable
  - Headword, hom, variants, pronunciation
- For every grammar block
  - The part-of-speech
- For every lexie
  - Government pattern for predicative lexies
  - Free text definition, domain, language levels
  - Examples, idioms
  - Link to the translation in the target language

Hypothesis on links

- Every translation link is bidirectional
- The frequency ranking of the links in their respective vocable can be different
- Example:
  - French “tabouret” $\xrightarrow{1}$ Japanese 椅子【いす】
  - Japanese 椅子 $\xrightarrow{1}$ French “chaise”
  - Japanese 椅子 $\xrightarrow{5}$ French “tabouret”

Linking process

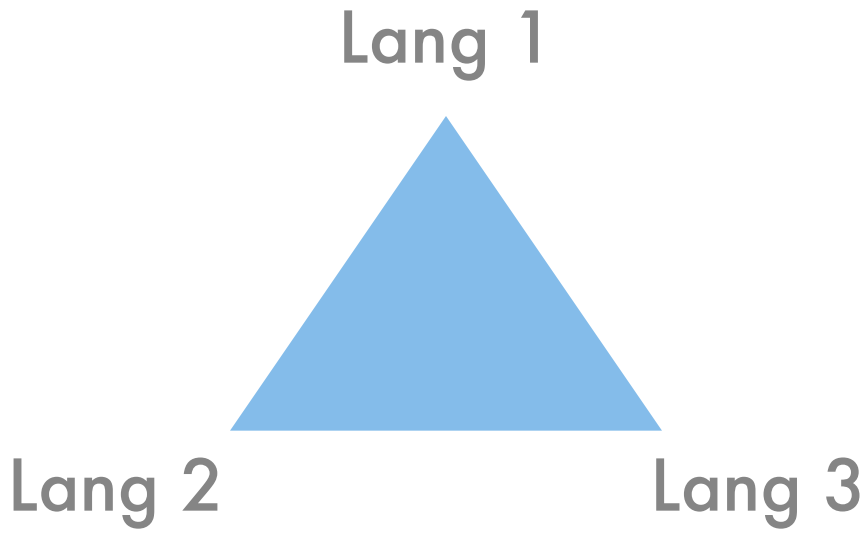
- Translation links done in a lexie block
- Not more than one link in a lexie block
- Link from vocable V1 to V2 =
  - Creation of a lexie L1 in V1
  - Creation of a lexie L2 in V2
  - Creation of a link L1->L2
  - Creation of a link L2->L1
    - The link is tagged “to be revised”

Writing process

- Anybody can contribute, logged or not
  - The data is publicly available right away
  - but it is marked “not revised”
- Trusted users can revise
  - The reviewer has to be a native speaker of the other language (than the contributor)
  - A user can be trusted if his/her contributions are always correct (after a certain period t)
- The contribution is then automatically validated

Which bilingual dict?

- Build 3 bilingual dicts in triangle
  - Useful for experimenting a pivot structure



Which language pair?

- French-Japanese
  - First aim of Papillon project (the roots)
- French-Malay or Vietnamese?
  - Already exist (FeM, FeV, VietDict)
- French-Thai
  - Exists only partially, seems a good solution
- Japanese-Thai
  - In order to close the triangle

Reusable data for fra-jpn

- Monolingual
  - French: Morphalou > 67,000 e. (entries)
  - Japanese: WaDokuJiTen > 214,000 e.
- Bilingual
  - Dico F-J > 10,000 e.; FJocean > 3,000 e.
  - Maniette 2,000 e.; Armement > 1,000 e.

Reusable data for fra-tha

- Monolingual
  - French: Morphalou > 67,000 e. (entries)
  - Thai: ?
- Bilingual
  - FeT dictionary

Reusable data for jpn-tha

- Monolingual
  - Japanese: WaDokuJiTen > 214,000 e.
  - Thai: ?
- Bilingual
  - SAIKAM

Conclusion

- My point of view after 4 years
  - Interest in Papillon project is decreasing
- ➡ Need a quick reaction
  - Lets launch contributive sub-projects
  - They will increase the motivation of the partners and the project visibility
- I need your agreement & collaboration
- Comment and react!