Papillon Lexical Database Project

Monolingual Dictionaries & Interlingual Links

Mathieu Mangeot

GETA/CLIPS IMAG

Grenoble, France

Mathieu.Mangeot@imag.fr

Plan

- Initiators & Partners of the Project
- Motivations & Goals of the Project
- General View & Architecture of the Database
- Structure of Monolingual Dictionaries
- Construction Methodology
  - Integration of Existing Resources
  - Adding of New Entries
  - Revision of New Entries
- Consultation of the Lexical Database
- Ongoing Work
- Conclusion & Contacts

Initiators & Partners

- Initiators:
  - Dr. Emmanuel Planas (GETA/CLIPS, France)
  - François Brown de Colstoun (French Embassy, Japan)
  - Dr. Mutsuko Tomokiyo (GETA/CLIPS, France)
- Partners:
  - **NII**: National Institute of Informatics (Tokyo, Japan)
  - GETA/CLIPS: Machine Translation (Grenoble, France)
- & numerous voluntary contributors

Motivations of the Project

- Lack of usage dictionaries (& in any case paying) French <-> Japanese USABLE by Francophones
- Lack of dictionaries for lingware
- Information not computerized
- Internet allows linguists, translators & researchers to collaborate easily
- Make data of the project available under open source license scheme

Goals of the Project: Production of Dictionaries

- For humans, in usual formats:
  - Internet consultation on-line
  - Paper edition
- For humans, thanks to databases:

  Direct help for editors, browsers or PDAs
- For machines:

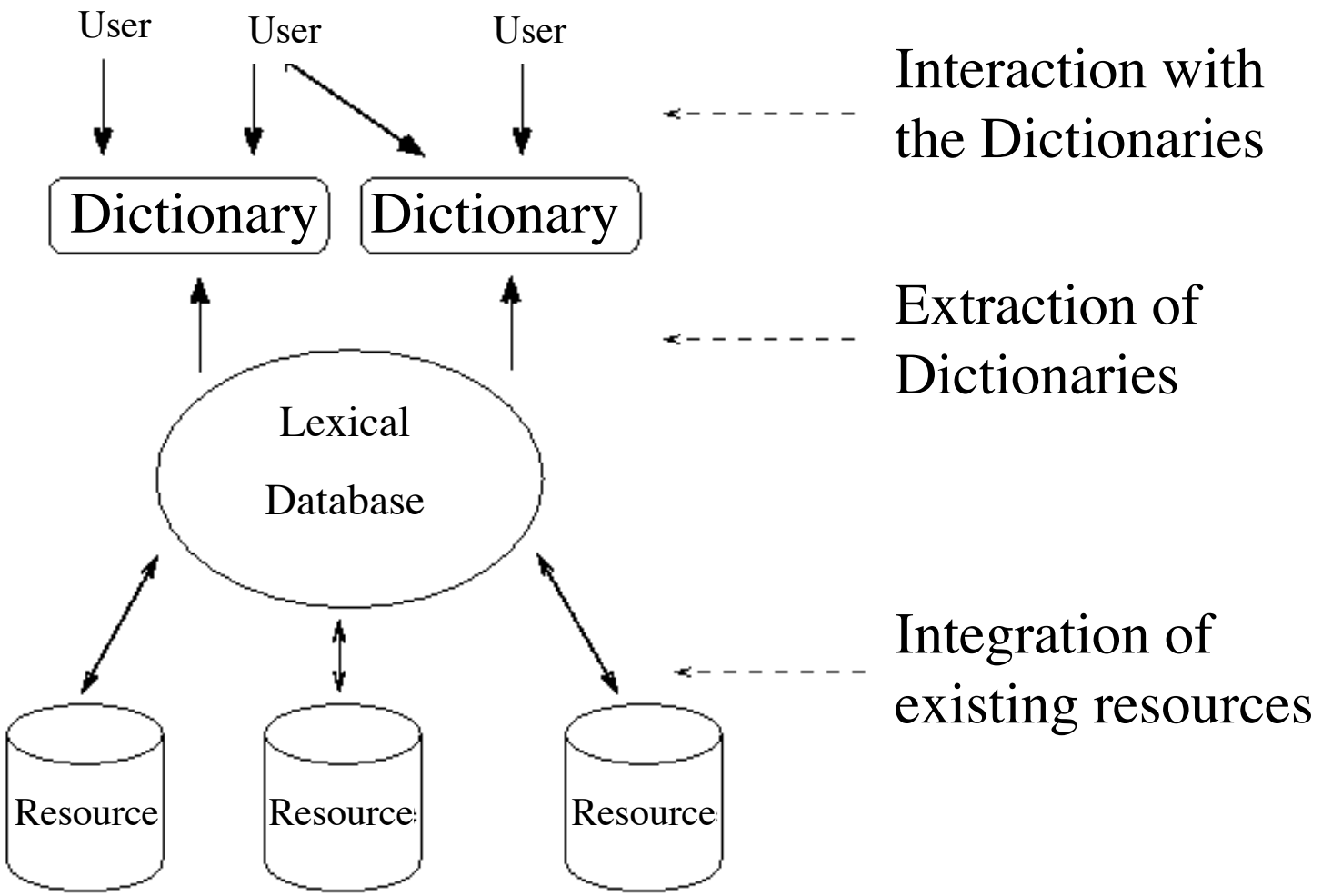
  Terminological resources for lingware
- For Science:

  Creation of multilingual dicos from monolingual ones

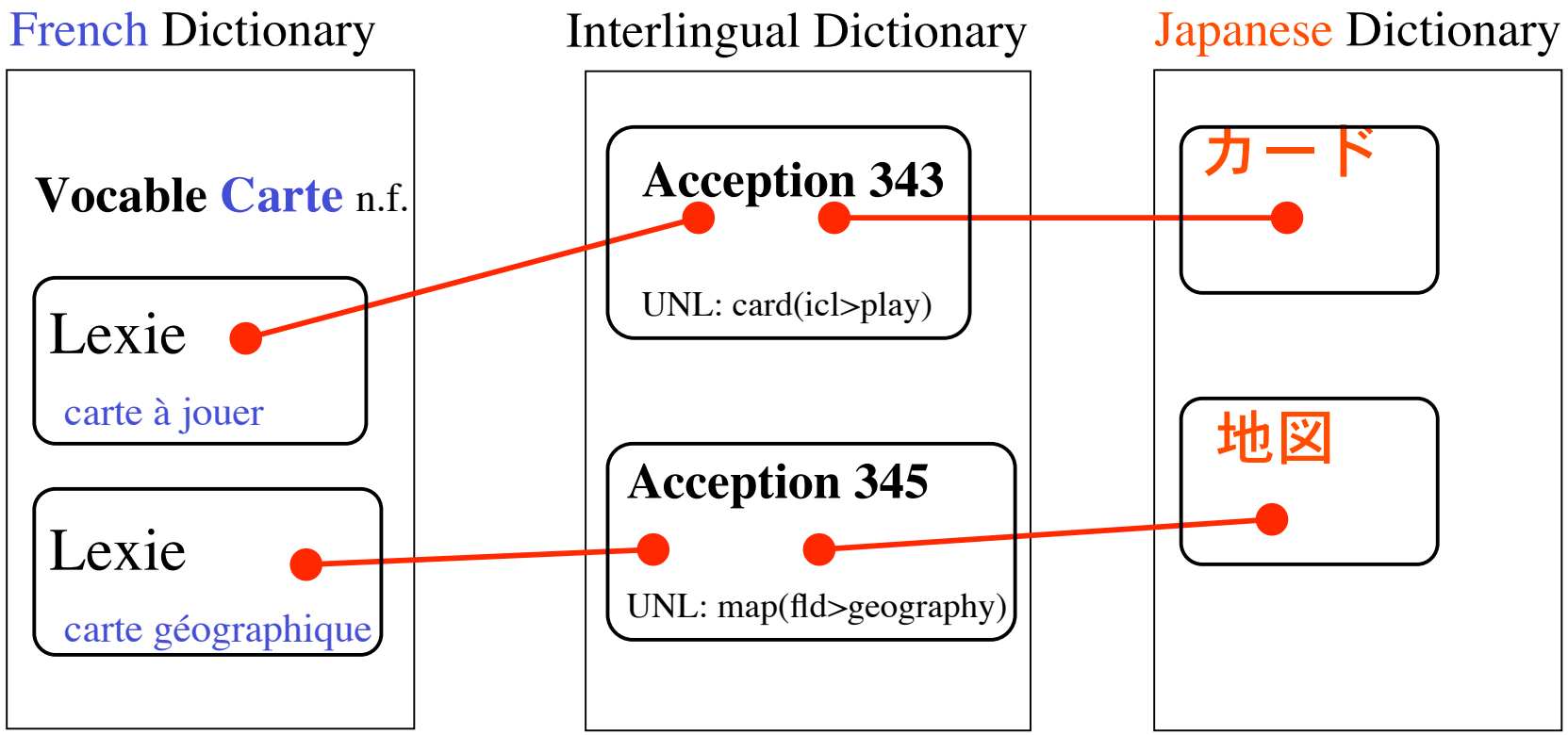
References & Previous Work

- Data:
  - FeM French->English-Malay - M. Lafourcade
    (Ass. Champollion/GETA, Grenoble; USM, Penang; DBP, KL)
  - JMDict Japanese->English - Jim Breen
    (Monash University, Clayton, Australia)
- Entry Logical Structure:
  - DEC, DiCo & LAF - I. Mel'cuk & A. Polguère
    (Université de Montréal, Montréal, Canada)
- Interlingual Databases:
  - PARAX - E. Blanc - (GETA/CLIPS)
  - SUBLIM - Ph.D. thesis of G. Sérasset - (GETA)
- Collaborative project:
  - SAIKAM - (NII & NECTEC)

General View of the Database



Internal Architecture of the Database



Architecture Derived from Dr. Gilles Sérasset's Ph.D. Thesis

Monolingual Dictionaries

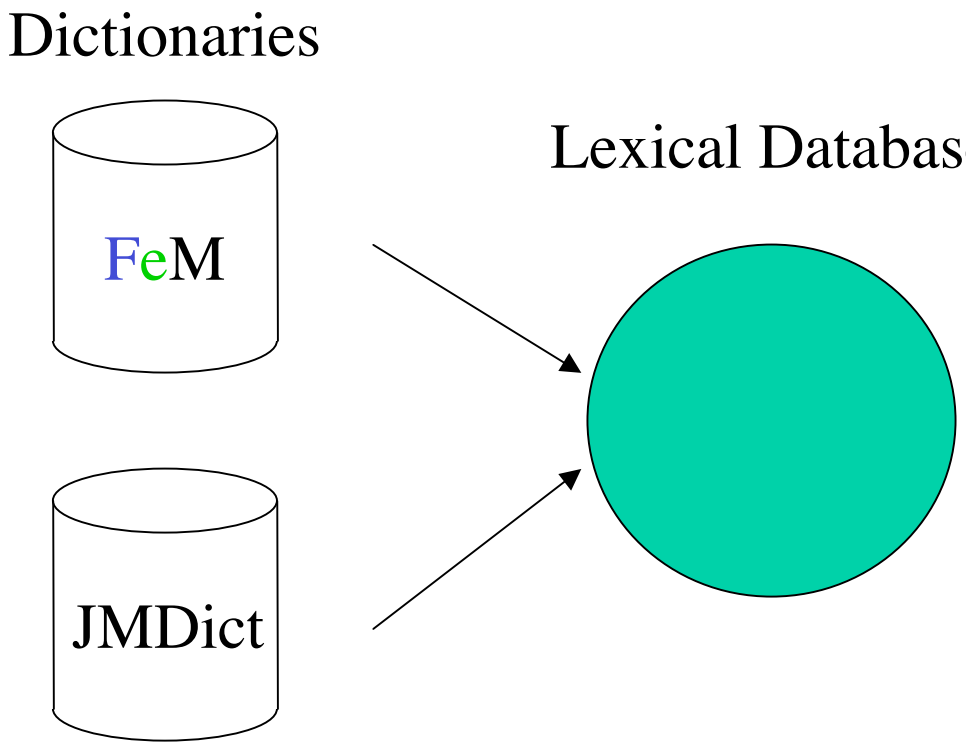
1. Name of the lexical unit: `MEURTRE`
2. Grammatical properties: `nom, masc`
3. Semantic Formula: `action de tuer: ~ PAR L'individu X DE L'individu Y`
4. Government pattern: `X = I = de N, A-poss Y = II = de N, A-poss`
5. (Quasi-)synonyms: `{QSyn} assassinat, homicide#1; crime`
6. Semantic derivations & collocations:
   - `{V0} tuer`
   - `{A0} meurtrier-adj /` *Nom pour X*/
   - `{S1} auteur [de ART Ø]` //meurtrier-n /*Nom pour Y*/
   - `{S2} victime [de ART Ø]` /*Très choquant*/
7. Examples: `La mésentente pourrait être le mobile du meurtre.`
8. Full Idioms:
   - `appel au meurtre`
   - `crier au meurtre`

Structure derived from Prof. Alain Polguère's Work on DiCo

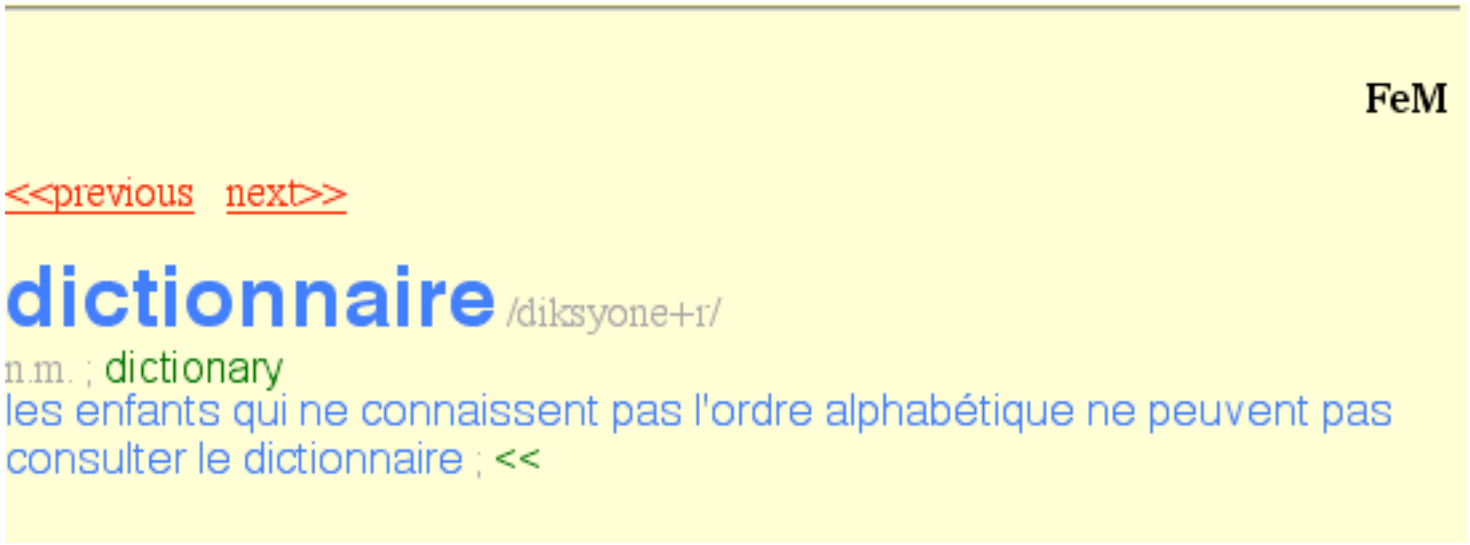
Construction Methodology

- Creation of the lexical soup
  - Integration of existing data
- Revision of the lexical soup
  - Revision of the links created automatically
- Creation of new data
  - The lexicographer writes monolingual entries
  - The translator edits interlingual links
- Revision of the data
  - The lexicologist reviews links & entries

Creation of the lexical soup



FeM: French->English



http://clips.imag.fr/geta/services/fem/

FeM structure

```
(:fem-entry
(:ENTRY "dictionnaire")
(:FRENCH_PRON "diksyone+r")
(:FRENCH_CAT "n.m.")
(:FRENCH_GLOSS                    "  u  n
texte")
(:ENGLISH_EQU "dictionary")
(:FRENCH_PHRASE  "les  enfants  qui  ne
connaissent  pas  l'ordre  alphabétique  ne
peuvent pas consulter le dictionnaire")
```

http://clips.imag.fr/geta/services/fem/

JMDict: Japanese->English

from Prof. Jim Breen, Monash University, Australia

EDICT

Search Key: 日本 (longest match found)

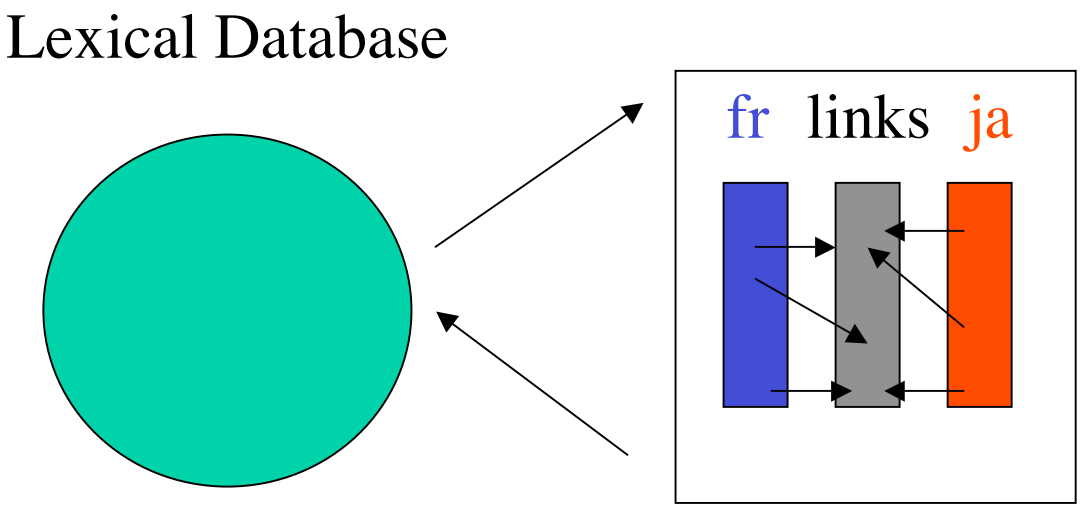
裏日本 【うらにっぽん】 Japan Sea coastal areas
裏日本 【うらにほん】 Japan Sea coastal areas
表日本 【おもてにほん】 Pacific side of Japan
日本 【にっぽん】 Japan
日本 【にほん】 Japan
日本猿 【にほんざる】 Japanese monkey; Japanese macaque
日本画 【にほんが】 Japanese paintings
日本海 【にほんかい】 Sea of Japan
日本海上自衛隊 【にほんかいじょうじえいたい】 Japan maritime self defense force
日本学者 【にほんがくしゃ】 Japanologist; Japan scholar

http://www.csse.monash.edu.au/~jwb/wwwjdic.html

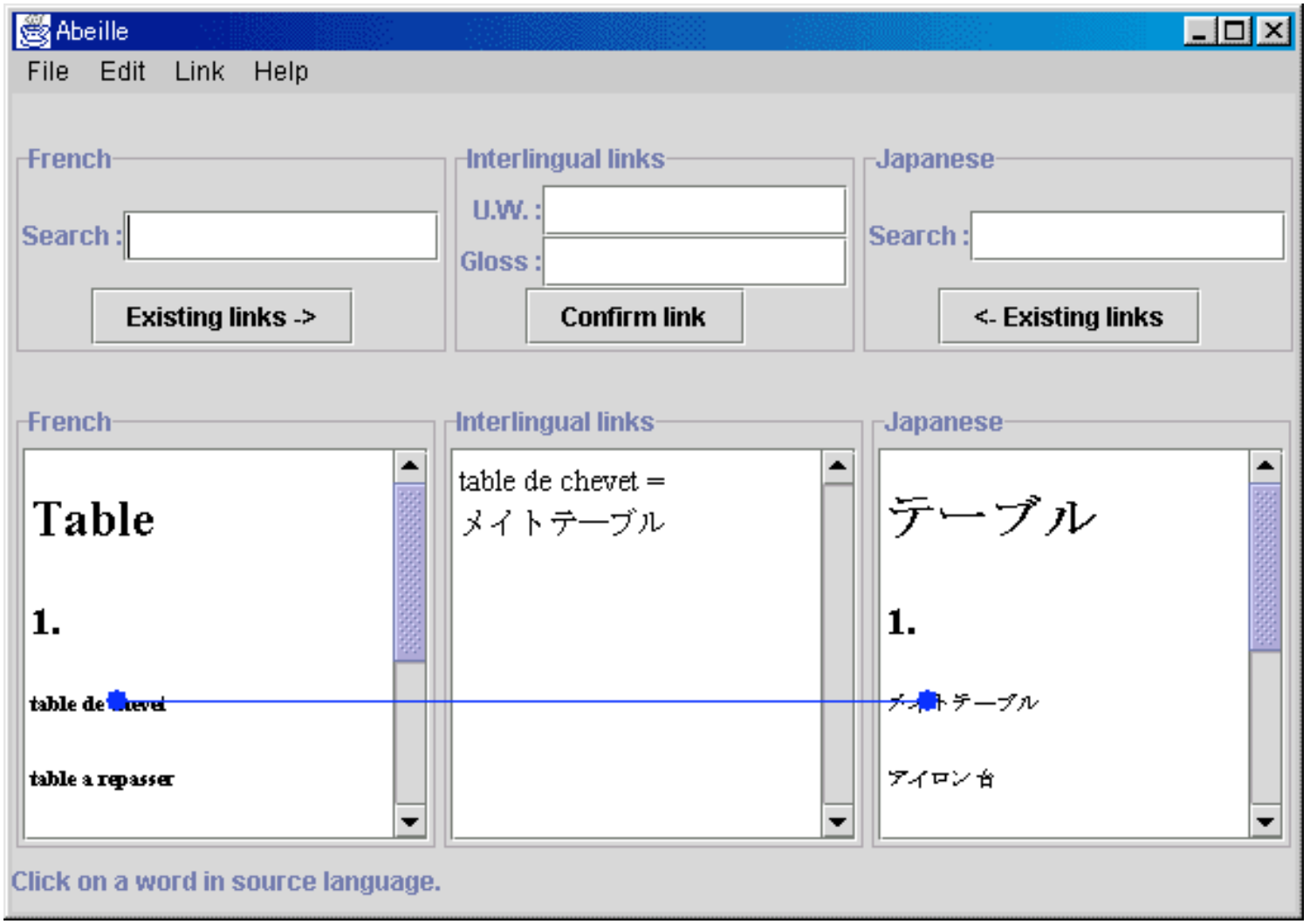
JMDict structure

```
<entry>
        <ent_seq>1582710</ent_seq>
        <k_ele>
                <keb>日本</keb>
                <ke_pri>jdd1</ke_pri>
        </k_ele>
        <r_ele>
                <reb>にほん</reb>
        </r_ele>
        <r_ele>
                <reb>にっぽん</reb>
                <re_pri>jdd1</re_pri>
        </r_ele>
        <sense>
                <gloss>Japan</gloss>
                <gloss g_lang="de">Japan</gloss>
        </sense>
</entry>
```

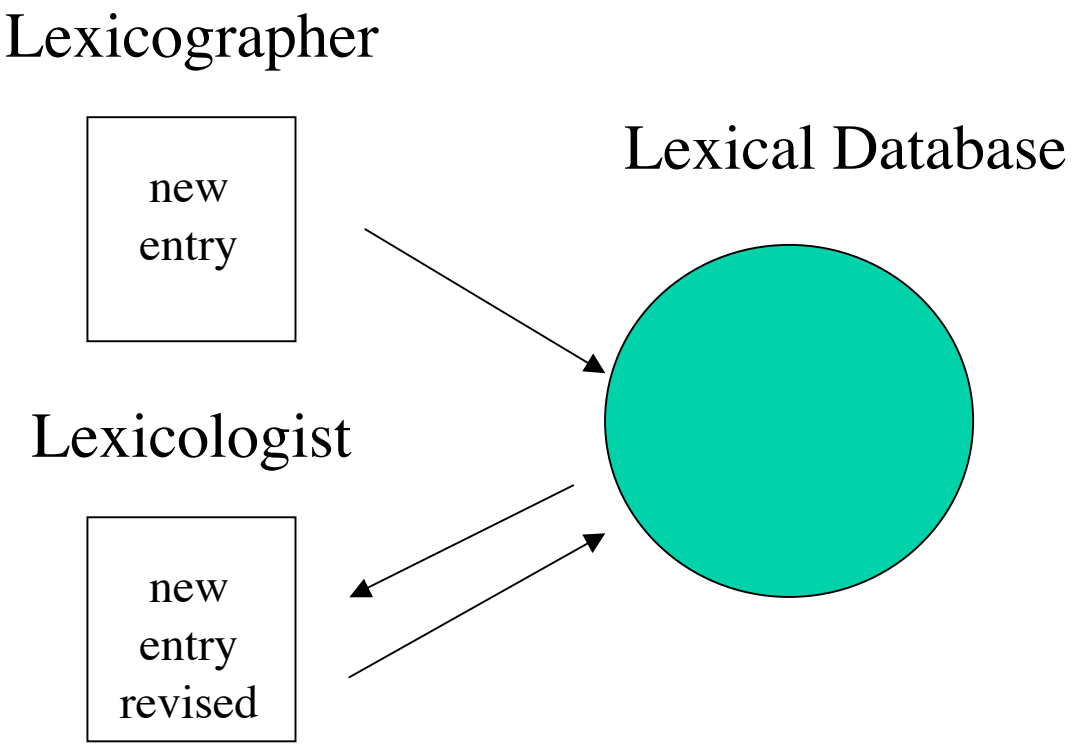
Revision of the Links



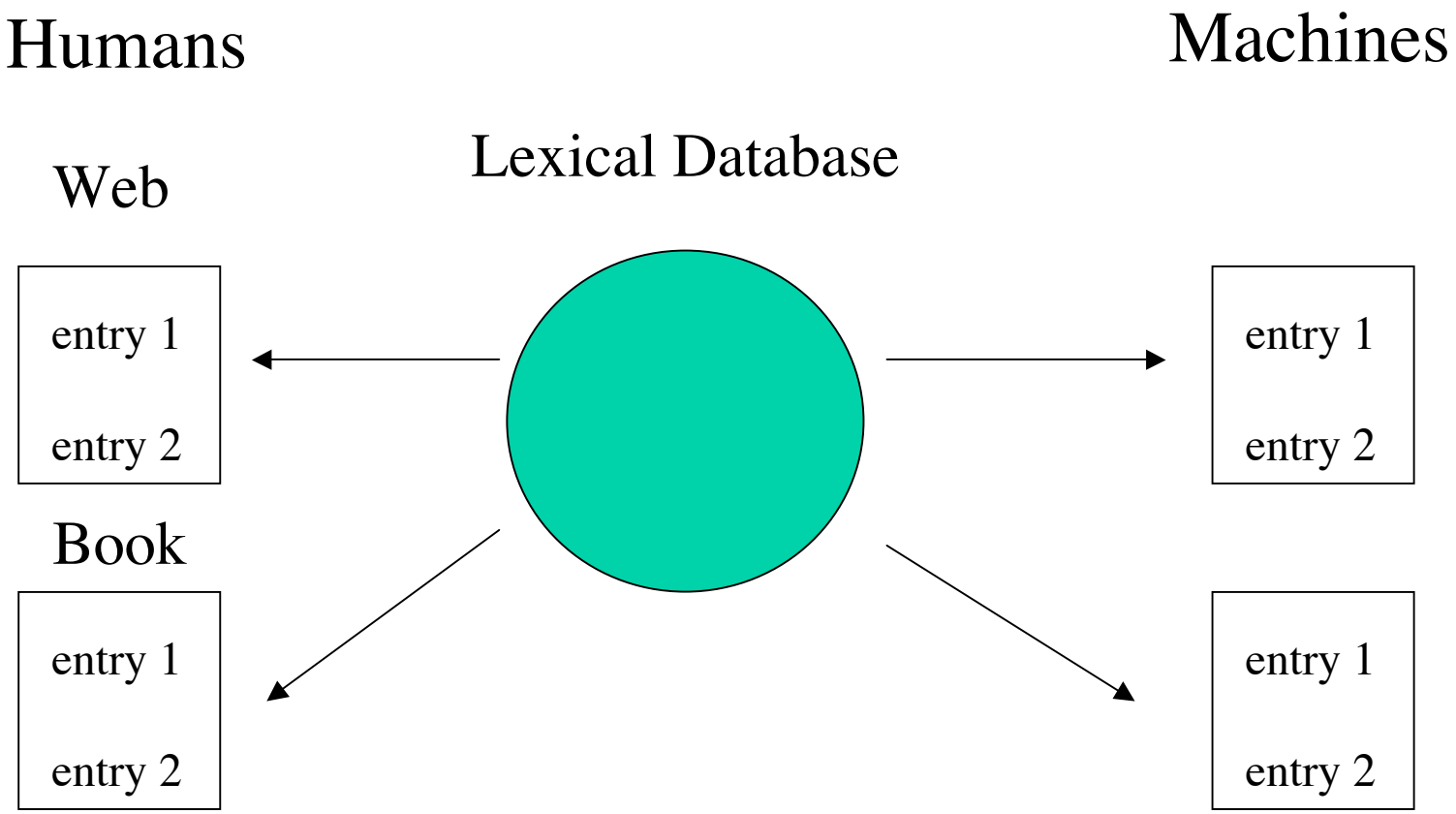
Revision Interface



Writing of New Entries



Consultation of the Database



Ongoing Work

- Ph.D. intern (Monthon) at NII, Tokyo
  - Preparation of the lexical soup with specific tools
- 4 months contract (M. Tomokiyo) (12/00-02/01)
  - Preliminary studies on linguistic content
- 2 years CNRS/JSPS grant at NII, Tokyo (10/2001–)
  - Management of the technical aspects of Papillon
  - Building of the server and CSCW tools
- Papillon 2001 workshop at Grenoble, France
  - July 2001, organized by GETA/CLIPS

Conclusion

- Advantages:
  - Easy integration of new languages
    - Ongoing discussions for Thai (KU & NECTEC) & Malay
  - Availability of the data with the open source license
  - Generation of multiple formats from the database
- Needs for the development of the project:
  - Centralized server & team of experts
  - Develop cooperative tools
  - Voluntary contributors !

Contacts

- Web Site: http://vulab.ias.unu.edu/papillon
- Responsible: Emmanuel Planas
  - mailto:Emmanuel.Planas@imag.fr
- Technical aspects: Mathieu Mangeot
  - mailto:Mathieu.Mangeot@imag.fr
- NII responsible: Frédéric Andrès
  - mailto:andres@nii.ac.jp

Construction Methodology

Commentaires	*Français*	*Japonais*	*anglais*	*UNL*
Idéal: résulat de la fusion	oui	oui	oui	oui
Idéal mais pas disponible	oui	oui	non	oui
Données du GETA	oui	non	oui	oui
FEM + JMdict	oui	oui	oui	non
Petites listes disponibles	oui	oui	non	non
FEM	oui	non	oui	non
JMdict	non	oui	oui	non
LADL-CNAM	oui	non	non	non
De zéro	non	non	non	non