Frameworks, Implementation & Open Problems for the Collaborative Building of a Multilingual Lexical Database

Mathieu Mangeot & Gilles Sérasset

NII, Tokyo, Japan

mangeot@nii.ac.jp

GETA-CLIPS, Grenoble, France

Gilles.Serasset@imag.fr

Outline

- Presentation of Papillon Project
- Macrostructure of the Dictionary
- Microstructure of the Entries
- Bootstrapping & Contribution Process
  - Limbo, Purgatory & Paradise
  - Bootstrapping with Conceptual Vectors
  - Contributions & Validation Process
- Lexico-Semantical Network
  - Monolingual with Lexical Functions
  - Multilingual with Axies (Interlingual Links)
- Conclusion & References

Motivations

- Initial Goal
  - Build a French-Japanese electronic dictionary for humans
- Lack of Information
  - Numerical Specifiers, kanji+kana+romaji
- Very Few Existing Resources
  - French-Japanese, Free, Electronic
- Construction Costs Too High
  - EDR English-Japanese Dictionary
  - 1200 human-year; 300 000 entries; price: 14,3 Mo ¥
- On Going Collaborative Construction Projects
  - Edict Japanese->English, SAIKAM Japanese-Thai

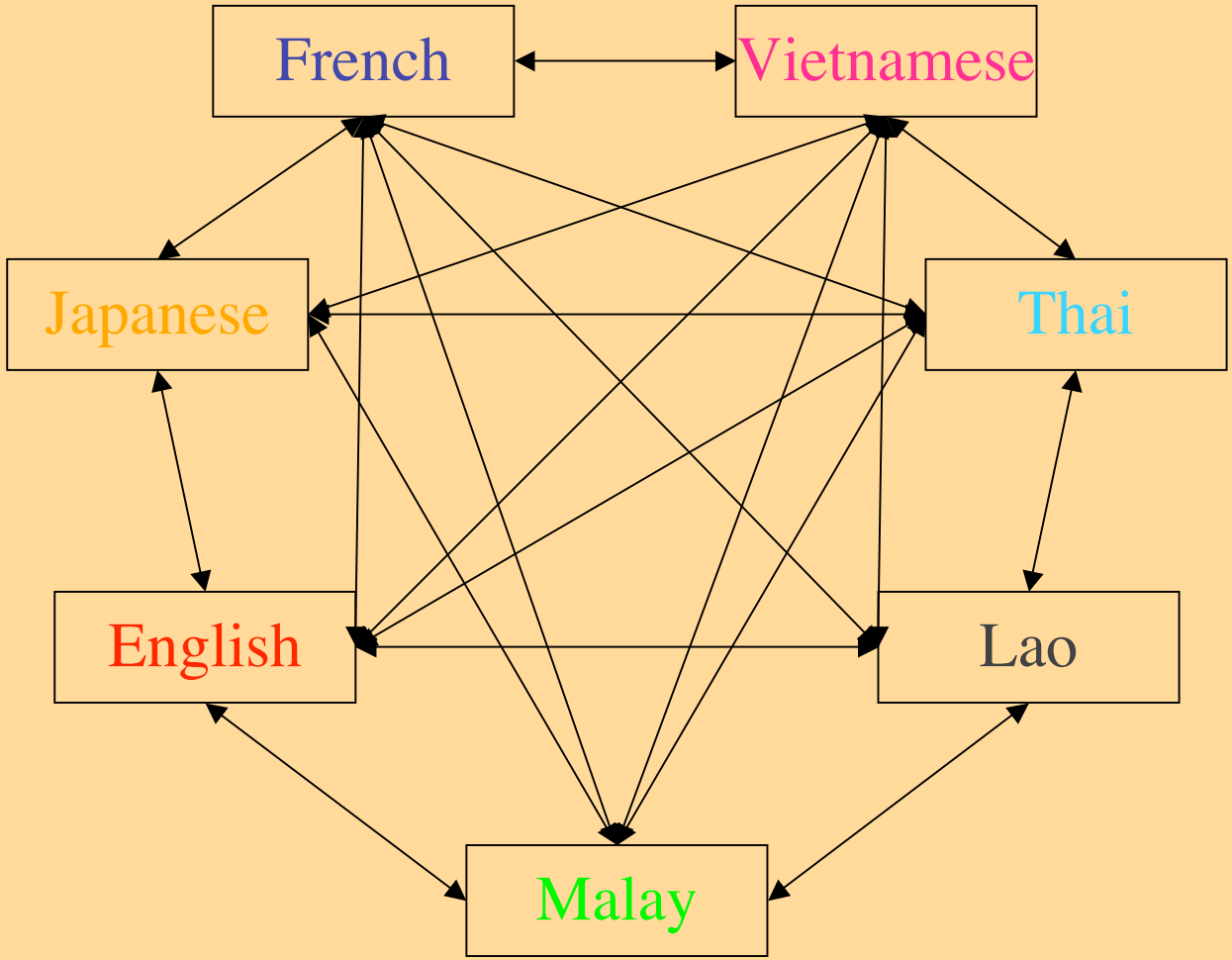
Extended Goals

- Build a More Complete Dictionary
  - Multilingual (English, French, German, Japanese, Lao , Malay, Thai, Vietnamese)
  - Multiusers (beginners, experts, applications)
- Community Development
  - LINUX Construction Paradigm
  - Voluntary Contributors
  - Mutualization of the Resources
  - User Preferences & Profiles

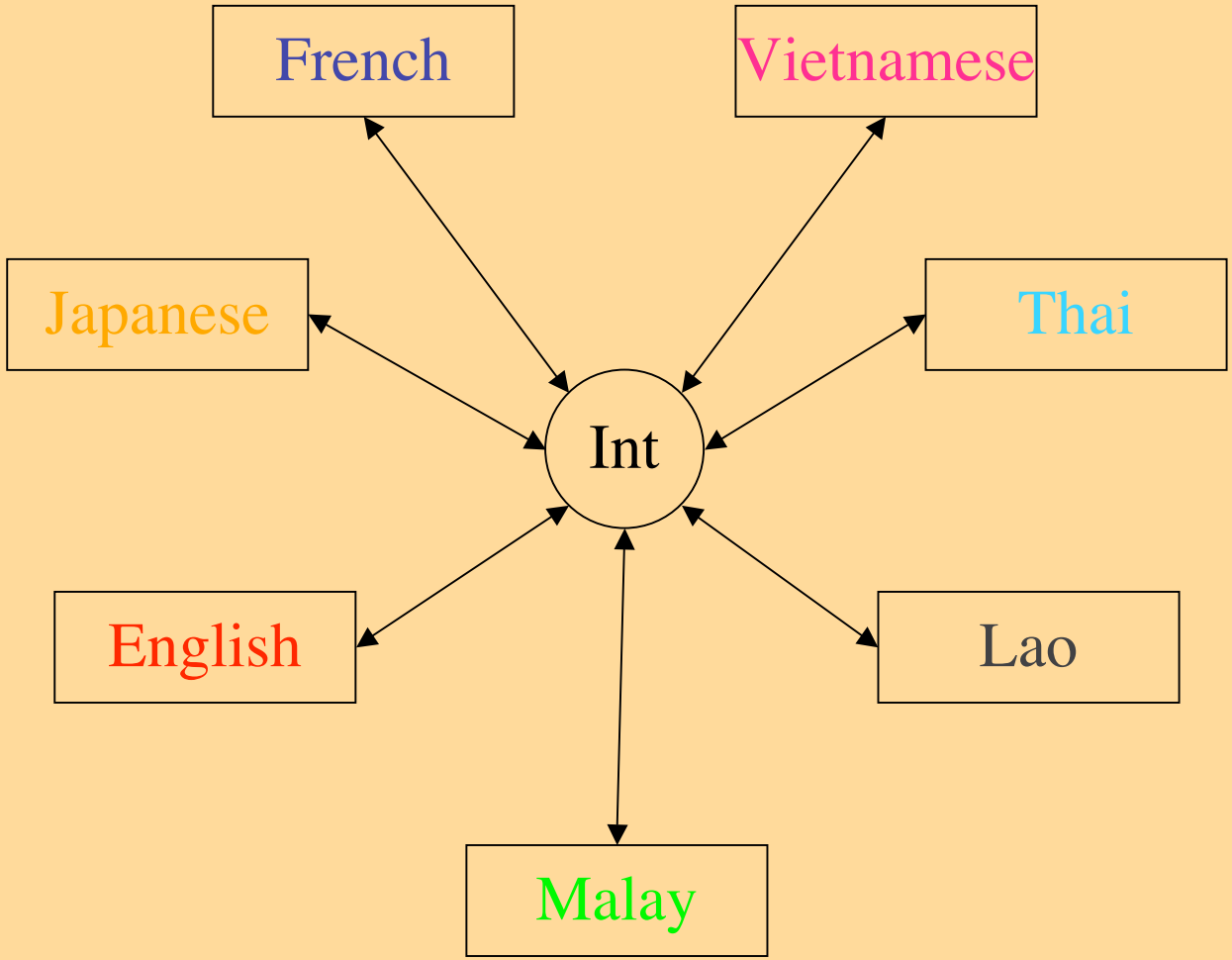
Outline

- Presentation of Papillon Project
- Macrostructure of the Dictionary
- Microstructure of the Entries
- Bootstrapping & Contribution Process
  - Limbo, Purgatory & Paradise
  - Bootstrapping with Conceptual Vectors
  - Contributions & Validation Process
- Lexico-Semantical Network
  - Monolingual with Lexical Functions
  - Multilingual with Axies (Interlingual Links)
- Conclusion & References

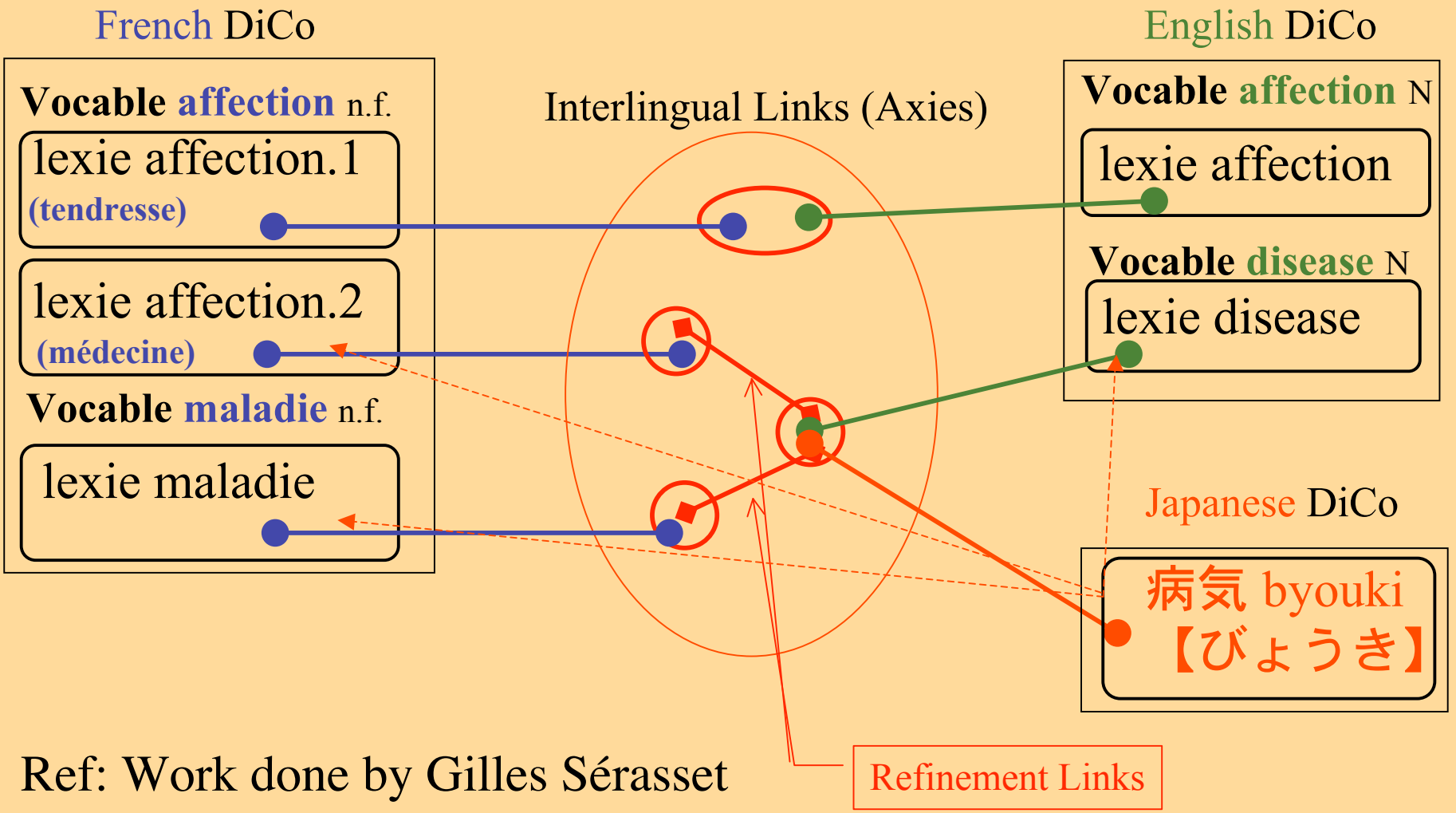
Bilingual Dictionaries



Pivot Dictionary



Detailed Pivot Structure



Ref: Work done by Gilles Sérasset

Outline

- Presentation of Papillon Project
- Macrostructure of the Dictionary
- Microstructure of the Entries
- Bootstrapping & Contribution Process
  - Limbo, Purgatory & Paradise
  - Bootstrapping with Conceptual Vectors
  - Contributions & Validation Process
- Lexico-Semantical Network
  - Monolingual with Lexical Functions
  - Multilingual with Axies (Interlingual Links)
- Conclusion & References

Combinatorial Lexicography

- From Meaning-Text Theory
  - Alain Polguère & Igor Mel'tchuk (U. de Montréal)
  - Gives the necessary information to go from an idea (the meaning) to its realisation in a given language (the text).
  - Existing Dictionaries: DEC, DiCo database & LAF
- Same Structure for Every Language
  - 56 Basic Lexical Functions

Papillon

French Lexie (DiCo Entry)

- Name of the Lexical Unit: `MEURTRE`
- Grammatical Properties: `nom, masc`
- Semantical Formula: `action de tuer: ~ PAR L'individu X DE L'individu Y`
- Government Pattern: `X =I = de N, A-poss Y= II = de N, A-poss`
- Lexical Functions:
  - `{QSyn} assassinat,homicide#1;crime` /*Quasi synonyms*/
  - `{Oper`$_1$`} accomplir, commettre, perpétrer [ART ~]; tremper [dans ART ~]` /*Causes that X does a M.*/
  - `{S`$_1$`} auteur [de ART Ø]//meurtrier-n` /*Name for X*/
  - `{S`$_2$`} victime [de ART Ø]` /*Name for Y*/
- Example: `La mésentente pourrait être le mobile du meurtre.`
- Idioms:
  - `_appel au meurtre_`
  - `_crier au meurtre_`

Japanese Lexie

- Name of the Lexical Unit:殺人 【さつじん】
- Reading: satsujin
- Grammatical Properties: 名詞 【めいし】
- Semantical Formula: どうさ: 人 Y の 人 X の ~
- Government Pattern: X = I = N, Y = II = N の
- Lexical Functions:
  - {QSyn} 殺戮【さつりく】, 殺害【さつがい】 /*Quasi synonyms*/
  - {$Oper_1$} [~ を] する;[~ を]犯す /*Causes that X does a M.*/
  - {S_1} 殺人者 【さつじんしゃ】, 殺人鬼 【さつじんき】 /*Name for X*/
  - {S_2} 被害者【ひがいしゃ】 /*Name for Y*/
- Example: 喧嘩【けんか】は殺人【さつじん】の動機【どうき】になり得【え】るだろう。
- Idioms:
  - _殺人剣 【さつじんけん】_
  - _嘱託殺人 【しくたくさつじん】_

Interlingual Links (Axies)

- Linking Monolingual Lexies
- Motivated by existing translation links.
  - Not like concepts
- Links to other Axies
  - Synonyms, Refinement, Generalizations
- Links to External References
  - To be independent from any existing theory
  - Wordnet synsets, NTT Semantic category,
  - ONTOS or LexiGuide ontologies,
  - UNL Uws & Graphs etc.

Papillon

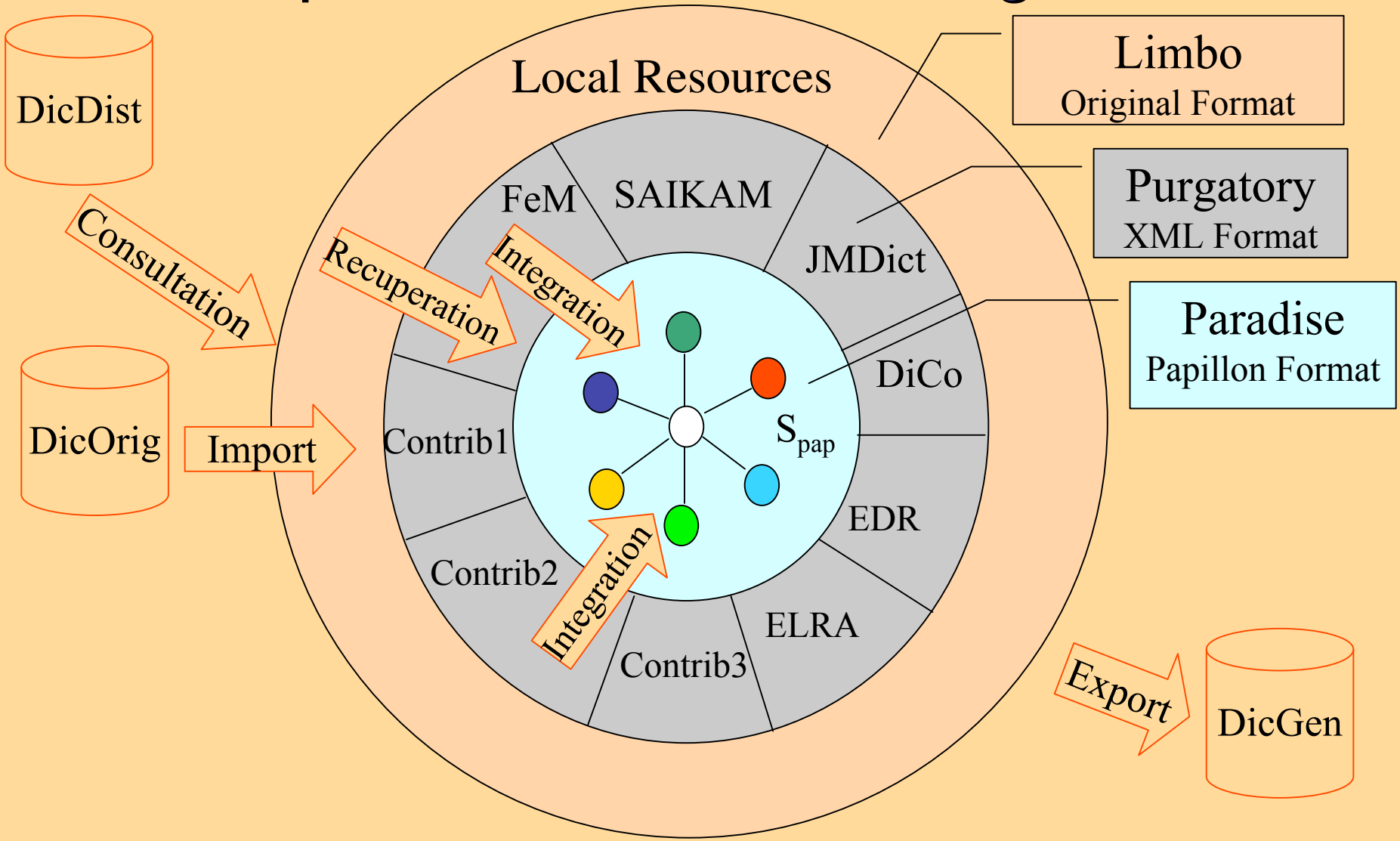
Structure of an Axie

- Unique ID: `a000023`
- Semantic Tag (entity, process, state, result): `process`
- Links to lexies: fra: `meurtre.1` eng: `murder.1` jpn: `satsujin.1`
- Links to other axies
  - synonym axies: `a000024 (assassination)`
  - generic axies: `a00002`
  - refined axies: `a000025`
- References to External Resources:
  - WordNet Synset: `00143589 unlawful premeditated killing of a human being`
  - UNL UW: `murder(icl>action,agt>human,obj>human)`
  - NTT Semantic Category
  - ONTOS Concept
  - LexiGuide concept

Outline

- Presentation of Papillon Project
- Macrostructure of the Dictionary
- Microstructure of the Entries
- Bootstrapping & Contribution Process
  - Limbo, Purgatory & Paradise
  - Bootstrapping with Conceptual Vectors
  - Contributions & Validation Process
- Lexico-Semantical Network
  - Monolingual with Lexical Functions
  - Multilingual with Axies (Interlingual Links)
- Conclusion & References

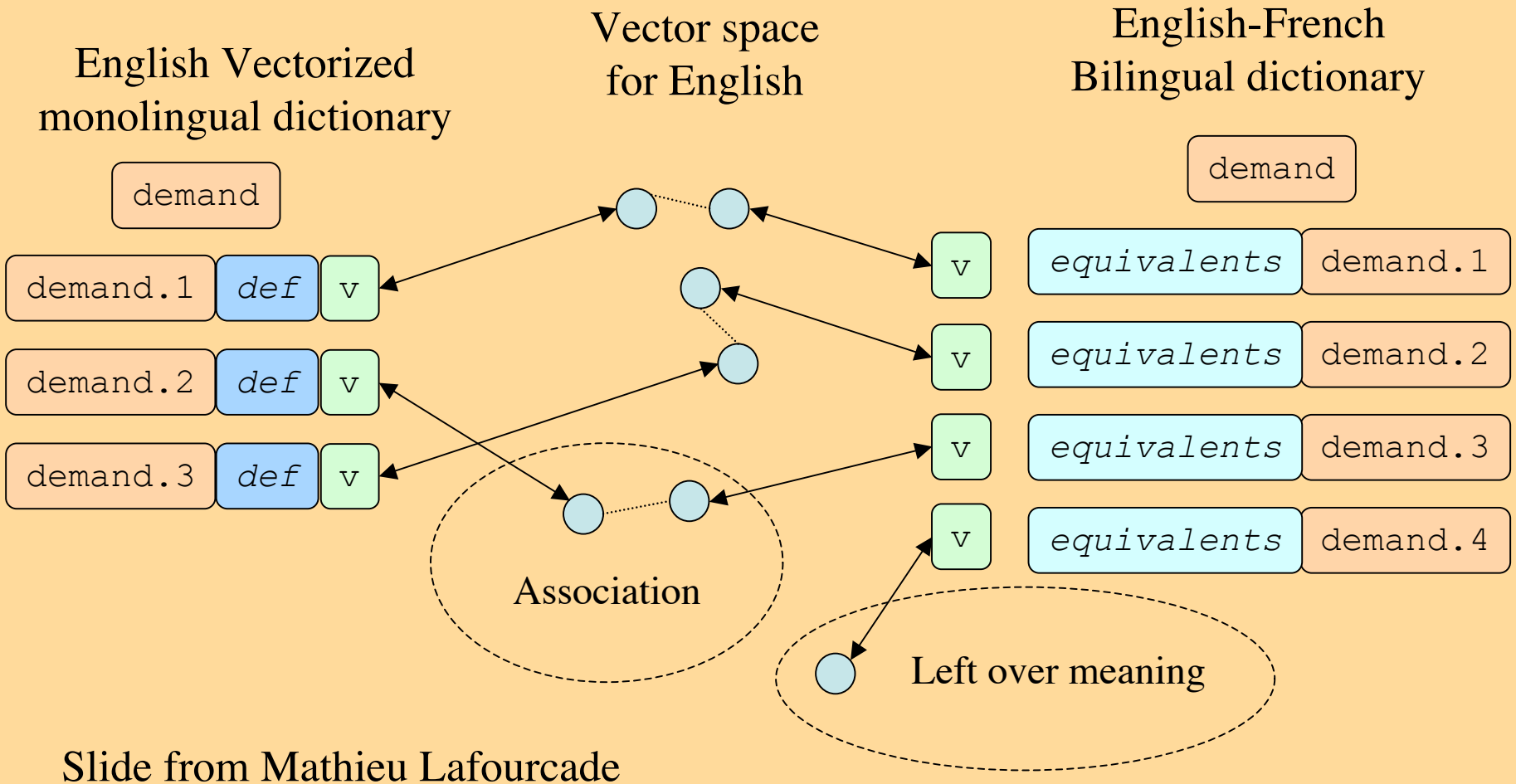
Preparation of the Existing Data



Introduction to Conceptual Vectors

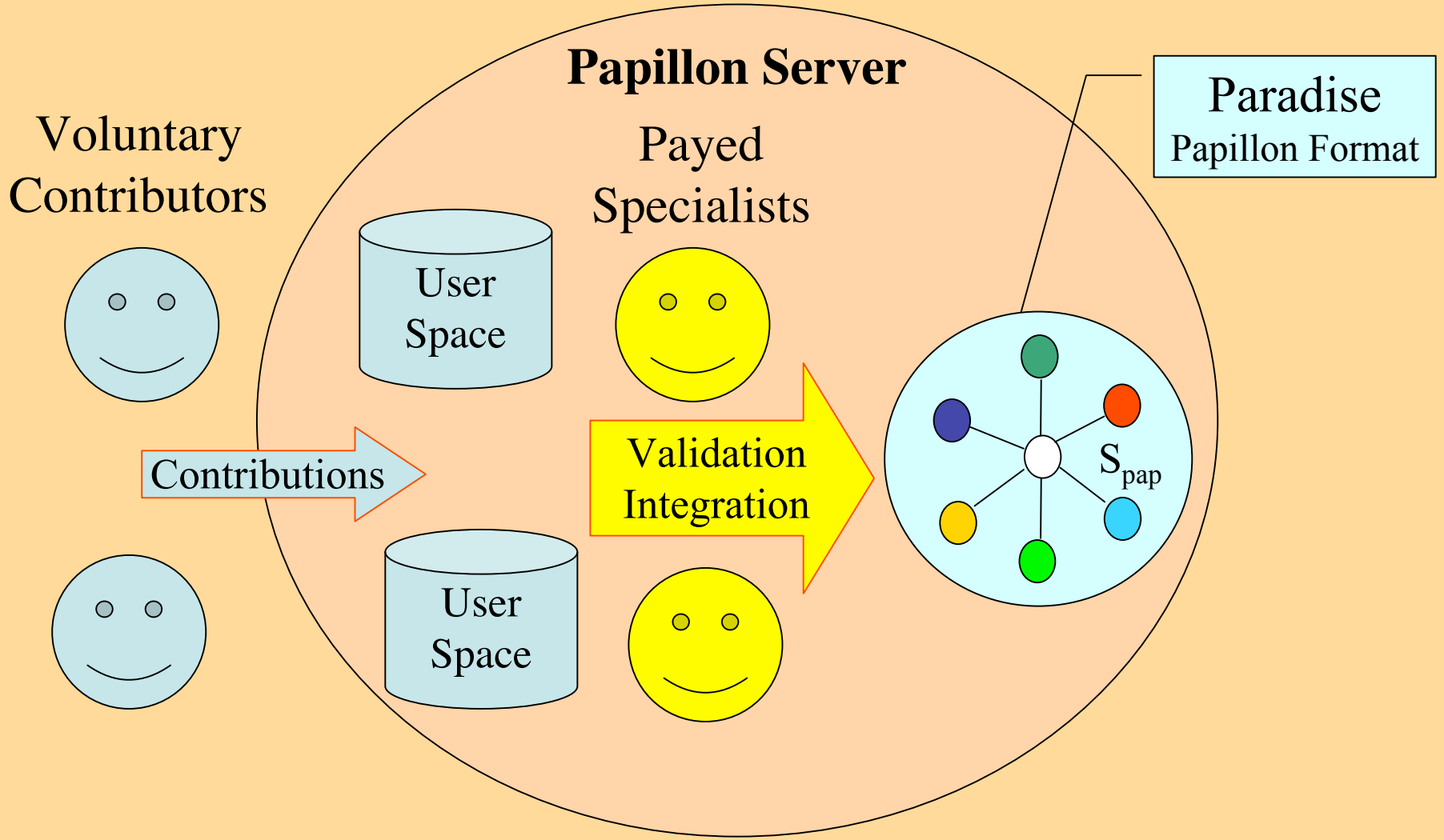
- An idea = a concept = a conceptual vector
- The vector space is of K dimensions
  - K = nb of concepts in a thesaurus hierarchy
  - Eg: for French, Thesaurus Larousse = 873 concepts
  - One independent vector space for each language
- Distance between 2 vectors = angular distance
  - $D_A(x, y) = \text{acos}(\text{sim}(x,y))$
  - $D_A(x, y) = \text{acos}(x.y/|x||y|)$
- Ref: Work done by Mathieu Lafourcade
  - http://www.lirmm.fr/~lafourca/

Linking Word Senses with Vectors



Slide from Mathieu Lafourcade

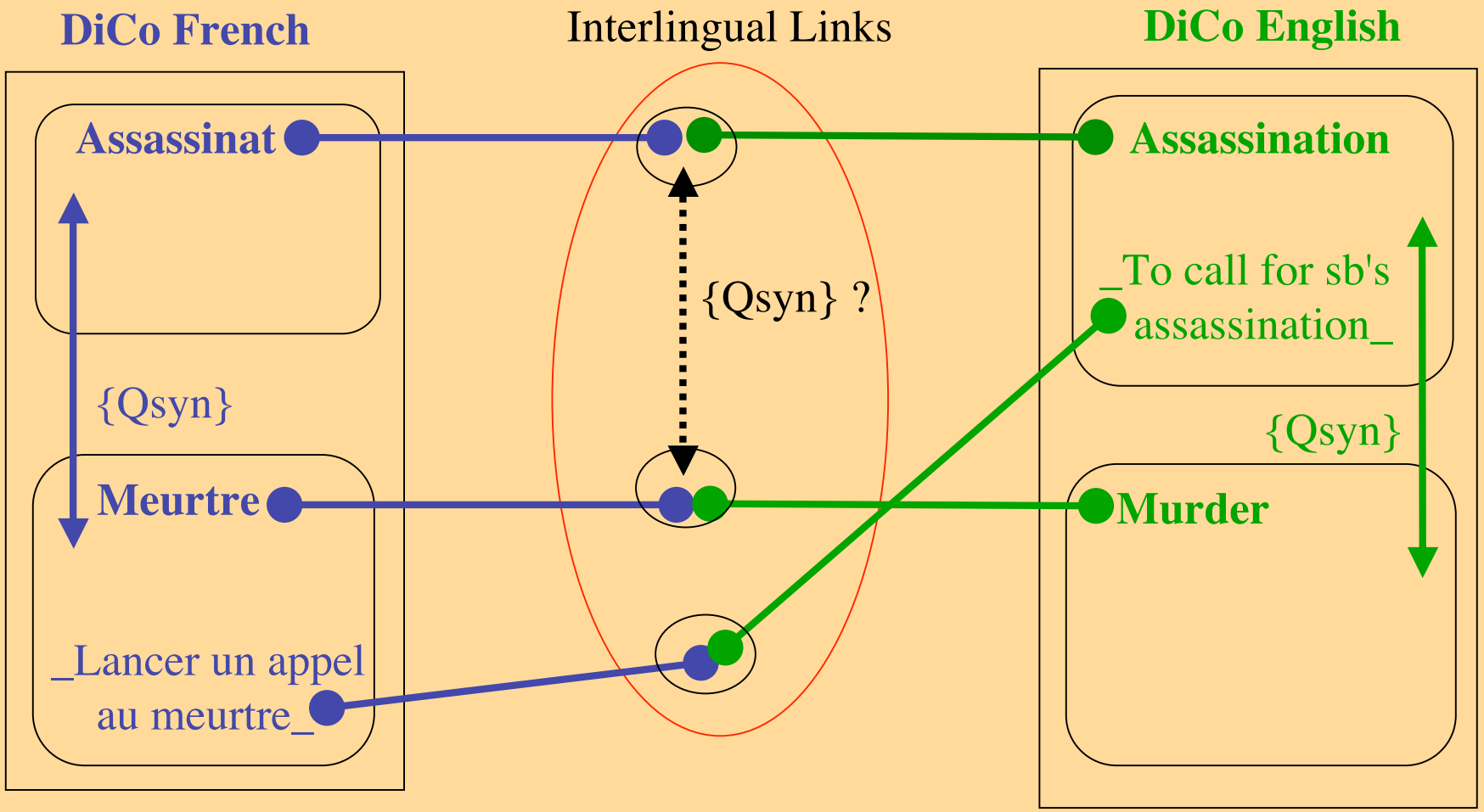
Contributions & Validation



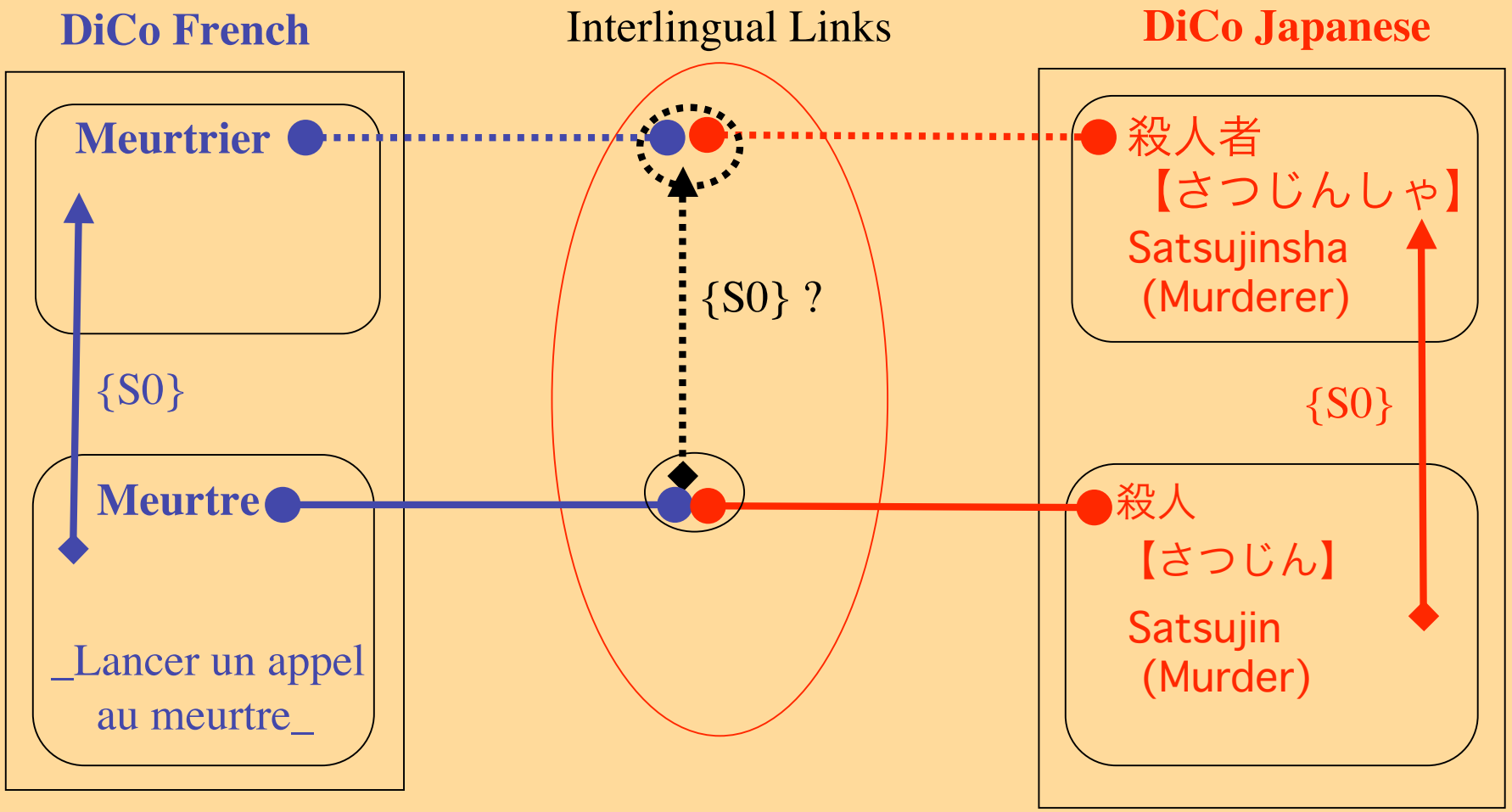
Outline

- Presentation of Papillon Project
- Macrostructure of the Dictionary
- Microstructure of the Entries
- Bootstrapping & Contribution Process
  - Limbo, Purgatory & Paradise
  - Bootstrapping with Conceptual Vectors
  - Contributions & Validation Process
- Lexico-Semantical Network
  - Monolingual with Lexical Functions
  - Multilingual with Axies (Interlingual Links)
- Conclusion & References

Lexico Semantical Multilingual Network (1)



Lexico Semantical Multilingual Network (2)



Outline

- Presentation of Papillon Project
- Macrostructure of the Dictionary
- Microstructure of the Entries
- Bootstrapping & Contribution Process
  - Limbo, Purgatory & Paradise
  - Bootstrapping with Conceptual Vectors
  - Contributions & Validation Process
- Lexico-Semantical Network
  - Monolingual with Lexical Functions
  - Multilingual with Axies (Interlingual Links)
- Conclusion & References

Conclusion

- Framework for Experimenting Networks
- Research Issues Remaining
  - Social issues: how to motivate people?
  - Contribution Interfaces
  - Checking Interfaces
- The Project Cannot Succeed without the Help of the Public People (Voluntary Contributors)

References & Contacts

- Web Site (information & consultation)
  - http://www.papillon-dictionary.org
- Steering Committee President
  - Gilles Sérasset Gilles.Serasset@imag.fr
- Technical Responsible in Japan
  - Mathieu Mangeot mangeot@nii.ac.jp