FACULTY
OF INFORMATICS
Masaryk University

A Hypergraph Model for Building Multilingual Dictionary Applications

Louis Lecailliez, Mathieu Mangeot

Long-Term Goal

A Learner's Multilingual Dictionary Application

- Help learning multiple languages at once
- Focused on east-Asian languages
- Rise linguistic awareness
- Offline use

Main Issue: Heterogonous Data

Vastly Diverging Micro-structures of Existing Material

- Sometimes full dictionaries

Chanh 柽. Citron. (Formé des S. A. *mộc* 木, arbre, et *chinh* 征, s'avancer.)

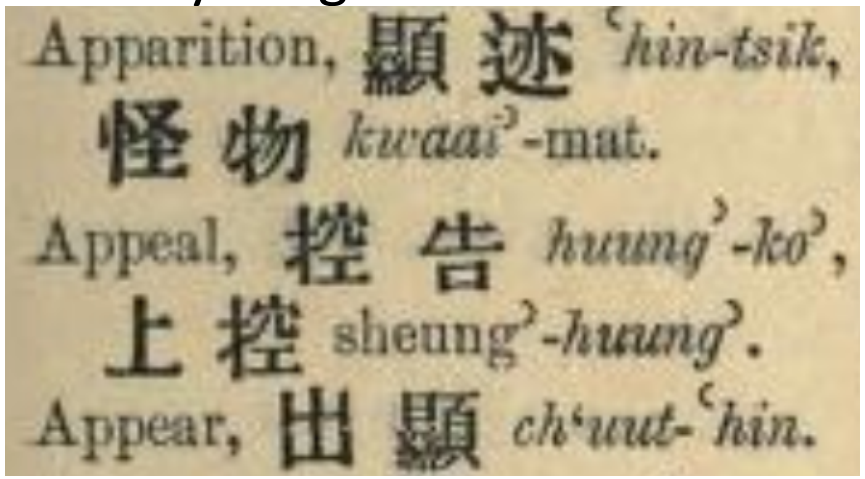
Cây chanh 核 ○, citronnier. — *Màu trái chanh* 牟 鞕 ○, couleur de citron. — *Nặn chanh* 攤 ○, exprimer le jus d'un citron. — *Nwớc chanh* 渚 ○, jus de citron. — *Nwớc mắm nặn chanh* 渚 鰳 攤 ○, espèce de saumure au jus de citron.

- May be as simple as a pair

掛 6312 467 kwà

掼 6346 vat 475 kwán

- Everything in between



Apparition, 顯迹 ʻhin-tsik, 怪物 kwaaiˀ-mat.

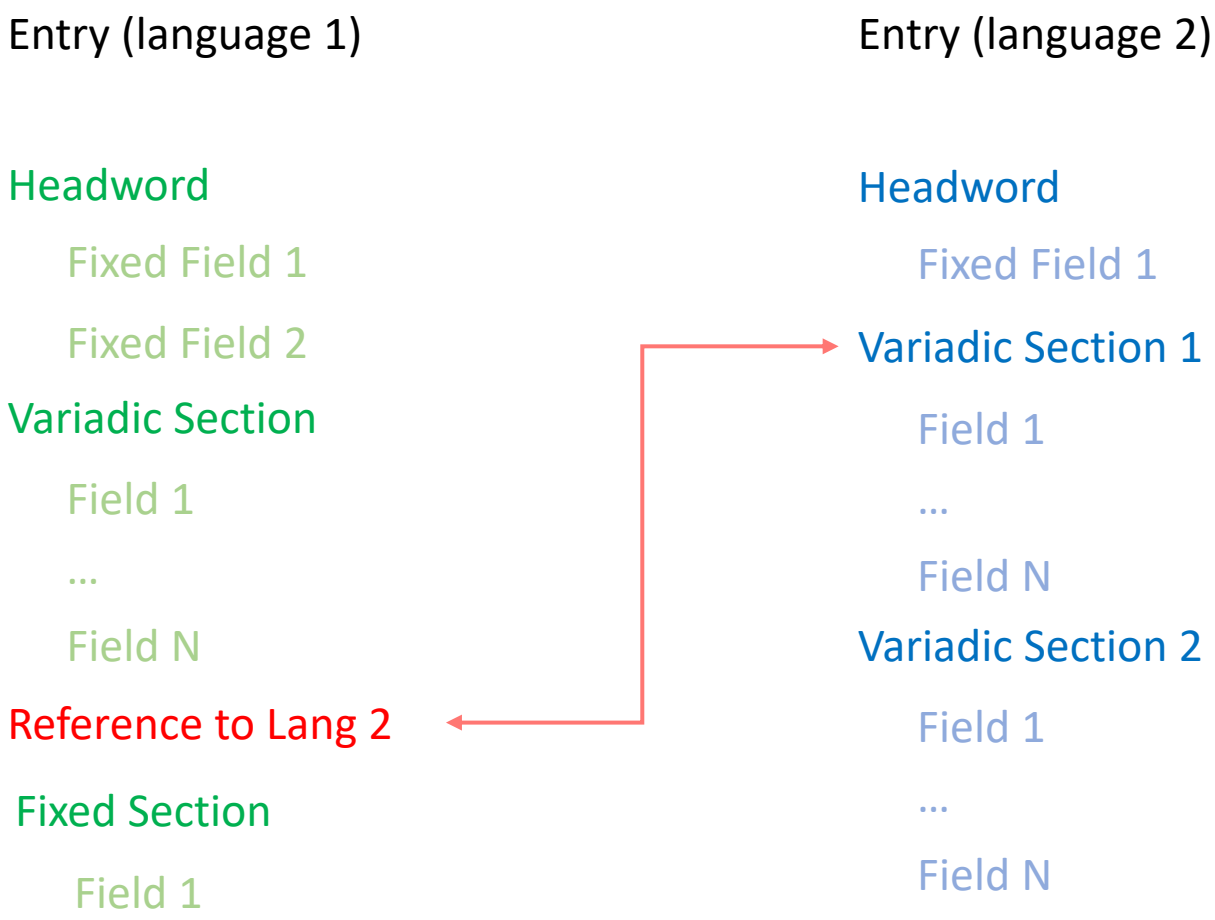
Appeal, 控告 huungˀ-koˀ, 上控 sheungˀ-huungˀ.

Appear, 出顯 chʻuut-ʻhin.

- Images from:
  - Bonet, 1899
  - MacIver, 1904
  - Charlmers, 1907

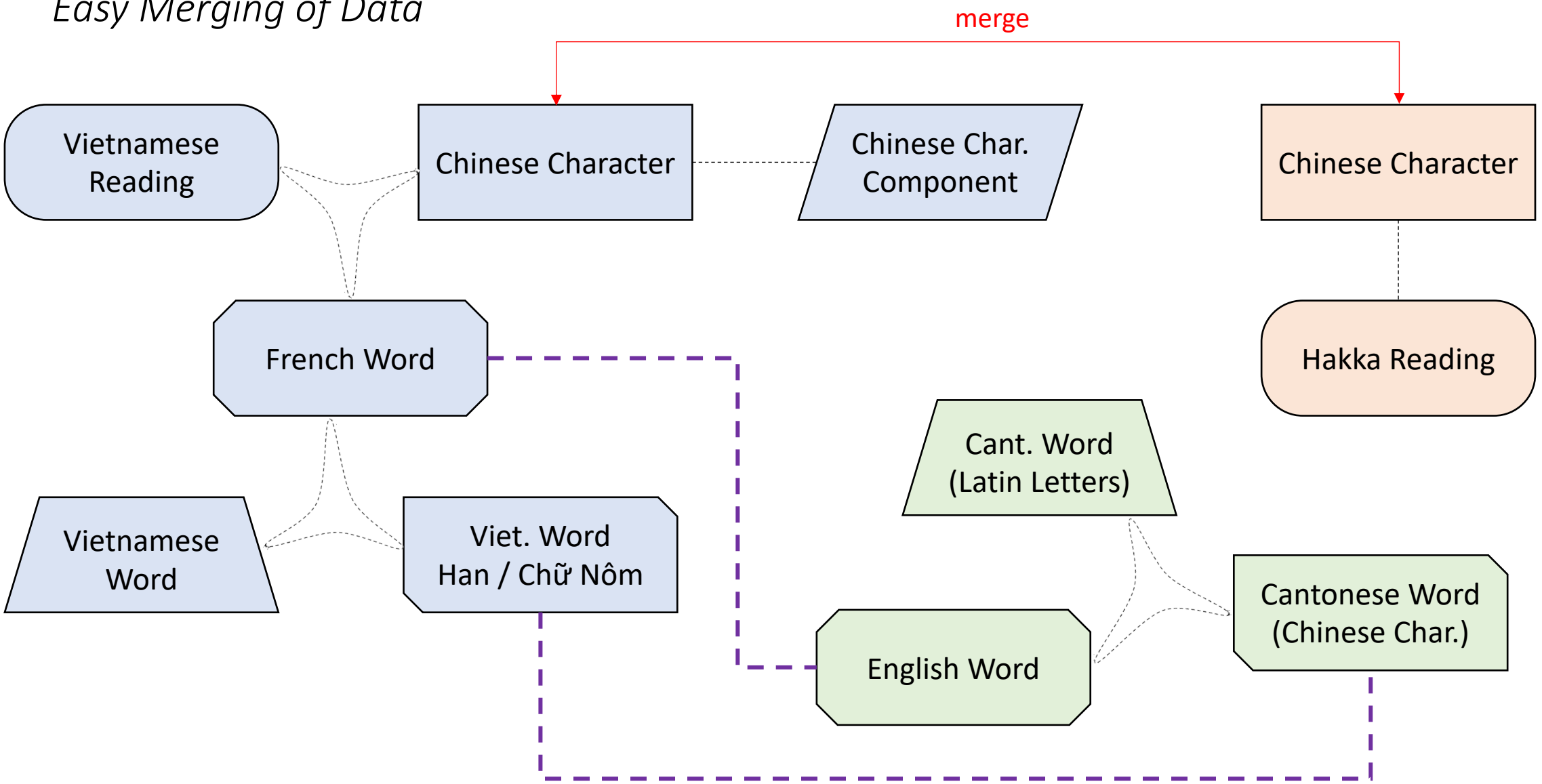
Issues with Traditional Modeling

Hierarchical Data & OOP Language & SQL Database Stack



One Solution: Graph

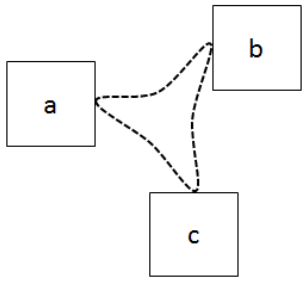
Easy Merging of Data



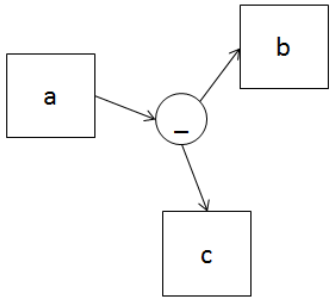
Issues with RDF

Existing Graph-like Data Model

- Complexity
- Not a general graph model
- Only features binary relations
  - Problem for handling Chinese characters



3-valant relationship



RDF modeling

- URI Scheme?
  - http://demo.com/
  - urn:uuid:f81d4fae-7dec-11d0-a765-00a0c91e6bf6
- Protocol?
  - http://louis.lecailliez.net
  - https://louis.lecailliez.net
- Base URI ending?
  - http://something.org/
  - http://something.org#

Proposed Graph Model

Type System

- Typing both nodes and edges
- No inheritance
- Use of UUID / GUID
  - Universally Unique Identifier
  - 128 bits number
  - f81d4fae-7dec-11d0-a765-00a0c91e6bf6

Property Name	Property Usage
Name	Human readable name
Identifier	Unique machine identifier (GUID)
Description	Human readable description
Object Kind	Vertex or Edge
is_oriented (edge only)	Boolean value
is_direct_content (vertex only)	Boolean value

Proposed Graph Model

Node Properties, is_direct_content Usage & Atomicity

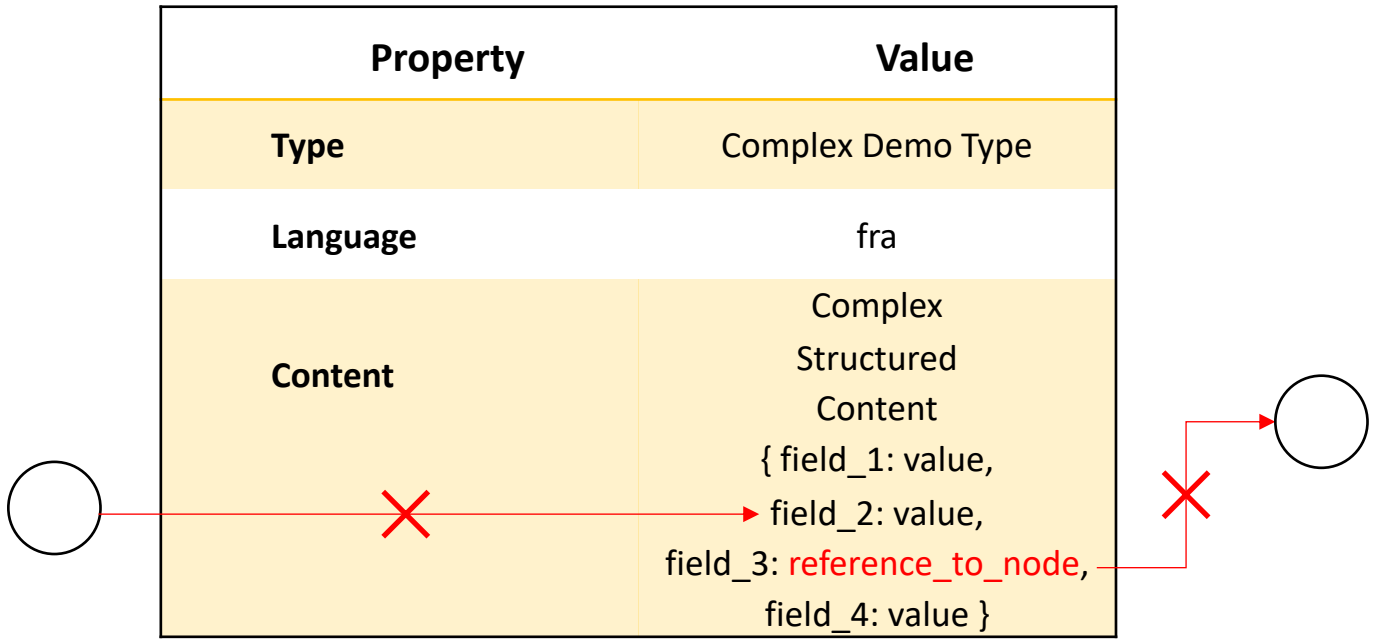
- Node Instance Properties

Property Name	Property Usage
Type	Type
Language	ISO 639-3 code
Content	Text

- is_direct_content usage

i_d_c: true	i_d_c: false
HiraganaWord	HW+DevoicedVowels
jpn	jpn
がくせい	がくせい/2

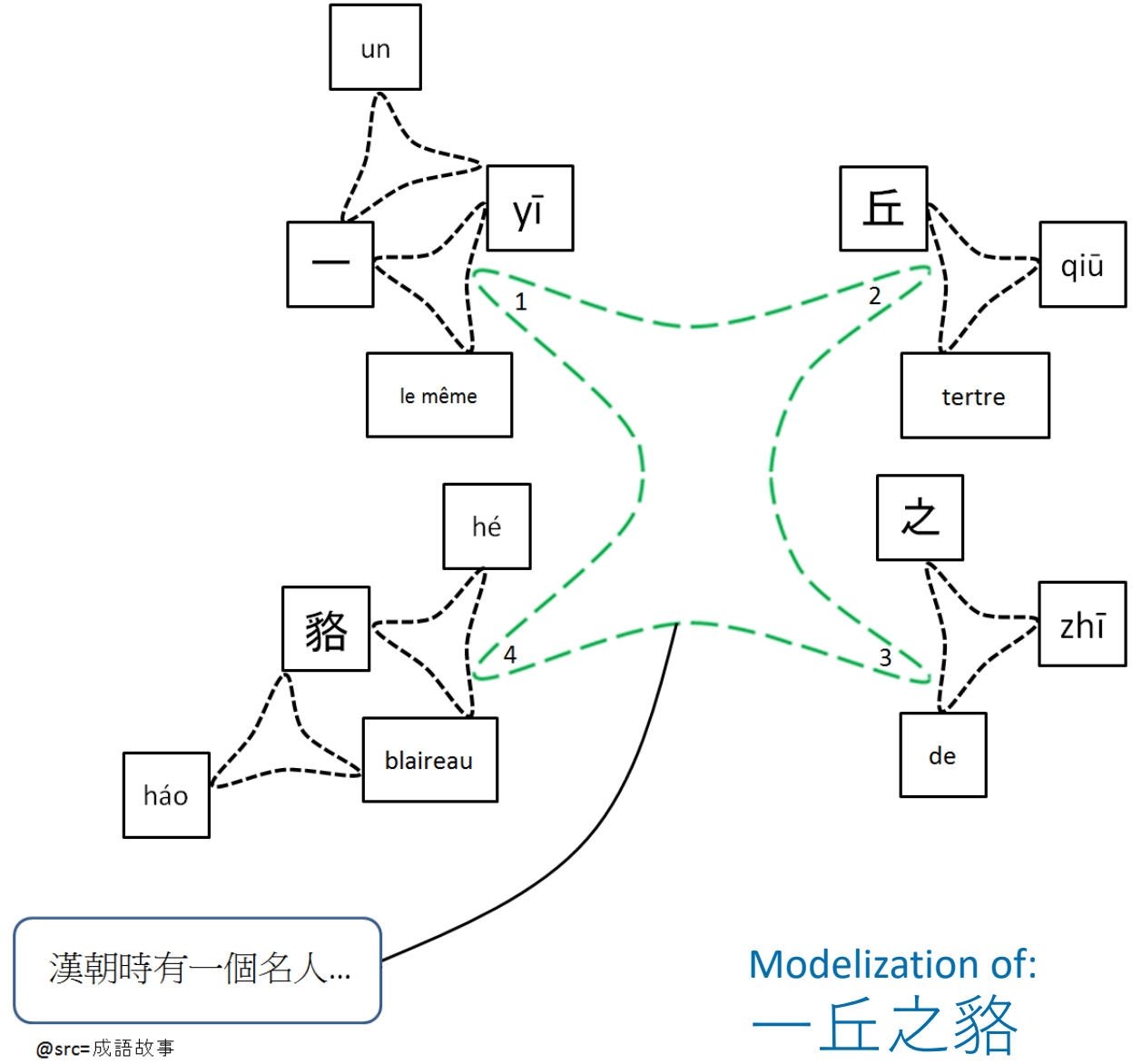
- Atomicity
  - No reference to the node content
  - No reference from a node content



Proposed Graph Model

Edge Types; Annotations

- Binary (oriented)
- Binary (non-oriented)
- Hyper-Edge
  - link more than two nodes (type I)
  - edge to edge (type II)
- All implemented in the same
  - uniform access from code
  - but hyper-edge II require more care
- Annotations
  - (namespace, key, value)
  - on both node or edge



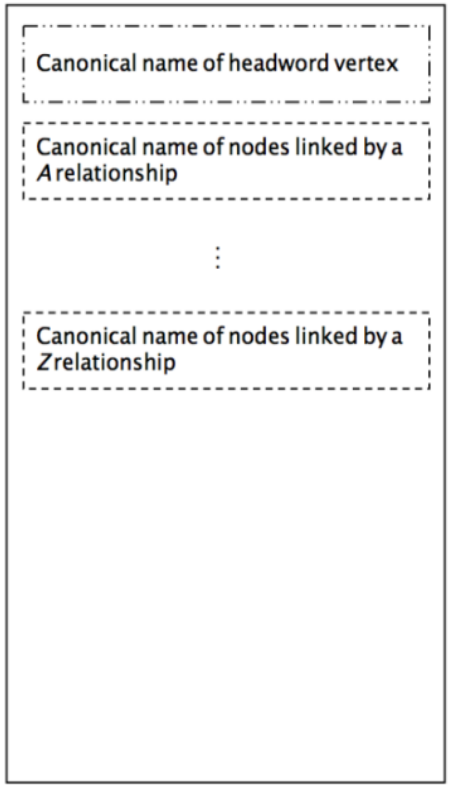
Modelization of:

一丘之貉

Page Display: Neighboring Nodes Display

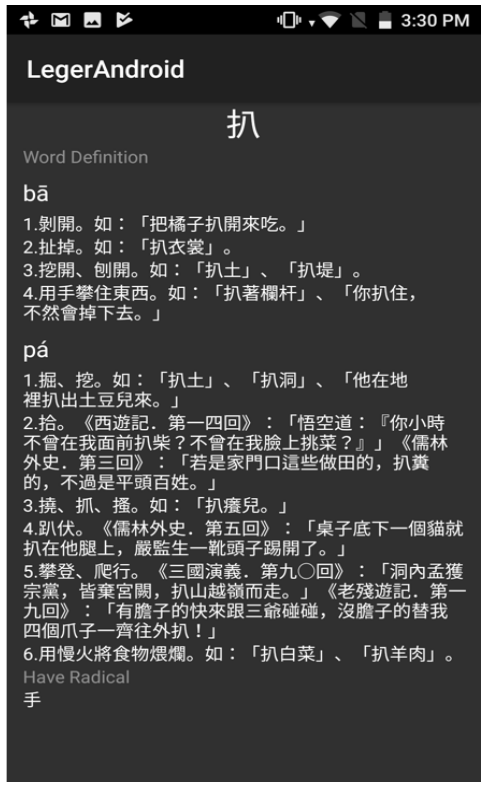
Nodes are grouped by relationship first

- Organization & Default Presenter
- Can be overridden



(a) Abstract Organization

(b) Concrete Example



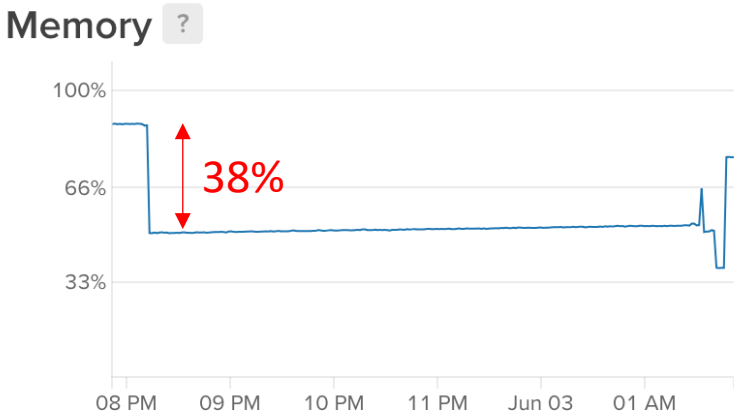
Discussion

Main Issue: Data Locality

- Issue in Data Access

	In Memory	On Storage
Startup	Slow	Fast
Data Access	Fast	Slow

- Not a problem on a server
  - E.g. IDS graph: 88,935 nodes & 195,896 edges
- Solution
  - finding a smart storage way in the literature

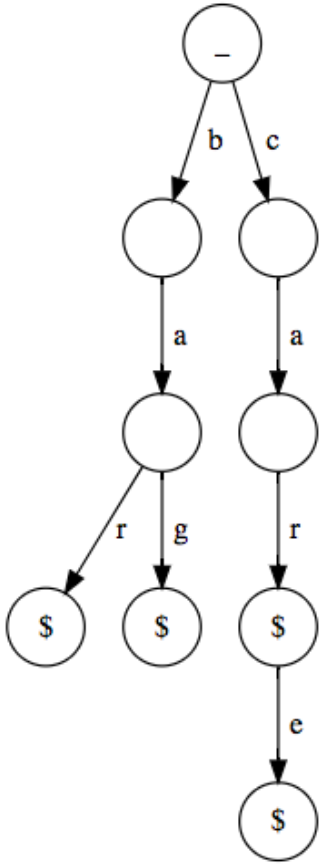


IDS Graph fits in 390 Mb of RAM

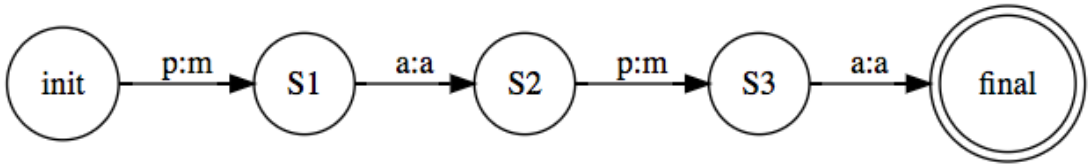
More Fun with Graphs

Implementation of Other Data-Structures

- Trie (Prefix Tree)
  - searching words by prefix



- Finite-State Transducer
  - rewriting strings



- Advantages
  - same data model
    - easy merge (if it makes sense)
  - same code base

Resulting Model

- Alleviate main RDF pain points:
  - does not take external dependencies (DNS, etc.)
  - uniform n-ary edge access
  - annotations mechanism
- Minimize application changes when data schema changes
  - may even not require modification at all in some cases

Implementation

Paper, Code & Prototype

- Paper missing on the USB Key!
  - louis.lecailliez.net
- Two Mobile Applications
  - two demo phones with me
  - ask me during the conference
- Code on GitHub
  - library only
  - one time code dump
  - BSD 3-Clause licensing

https://github.com/titanix/HyperGraph_Public

FACULTY
OF INFORMATICS
Masaryk University

Thanks for your attention

References

& Links

- Dictionaries:
  - Bonet, J. (1899). *Dictionnaire annamite-français: (langue officielle et langue vulgaire)* (Vol. 1). Imprimerie nationale, E. Leroux.
    - https://archive.org/details/dictionnaireanna01bone
  - Maclver, D. (1904). A Hakka Index to the Chinese-English Dictionary of Herbert A. Giles and to the Syllabic Dictionary of Chinese of S. Wells Williams.
    - https://archive.org/details/cu31924023344587
  - Chalmers, J., & Dealy, T. K. (1907). *English and Cantonese Dictionary* (Vol. 2). Kelly & Walsh, Limited.
    - https://archive.org/details/englishcantonese00chaluoft