Building a network of knowledge on sinograms

Pierre Magistry, Yoann Goudin,
Ilaine Wang and Guillaume Lechien
contact: pierre@magistry.fr

July, 2nd 2015
28èmes Journées de Linguistique d'Asie Orientale, Paris

The Team

- Pierre Magistry
  Ph.D. in Computational Linguistics,
  Freelance consultant in NLP between two postdocs
- Ilaine Wang
  Ph.D. candidate in Computational Linguistics
  MoDyCo, Univ. Paris X - Nanterre
- Yoann Goudin
  Ph.D. candidate in didactics, CERLOM, INALCO
- Guillaume Lechien M.A. in Computational Linguistics, works as a software engineer.

Evolution in how information can be stored

Law of the Instrument

"I suppose it is tempting, if the only tool you have is a hammer, to treat everything as if it were a nail."

One way to see it is as a matter of dimensionality in how we can access the information

- texts have 1 dimension
- tables have 2 dimensions
- graphs have *n* dimensions

- 爾雅
- 説文解字
- 韻鏡 (?)

About the Information Related to Sinograms

- multilingual
- graphical (decomposition on various level)
- lexical (composition on various level)
- phonological (at a certain level of decomposition)
- etymological (even when sinograms went out of use)
- available online in astonishing quantity

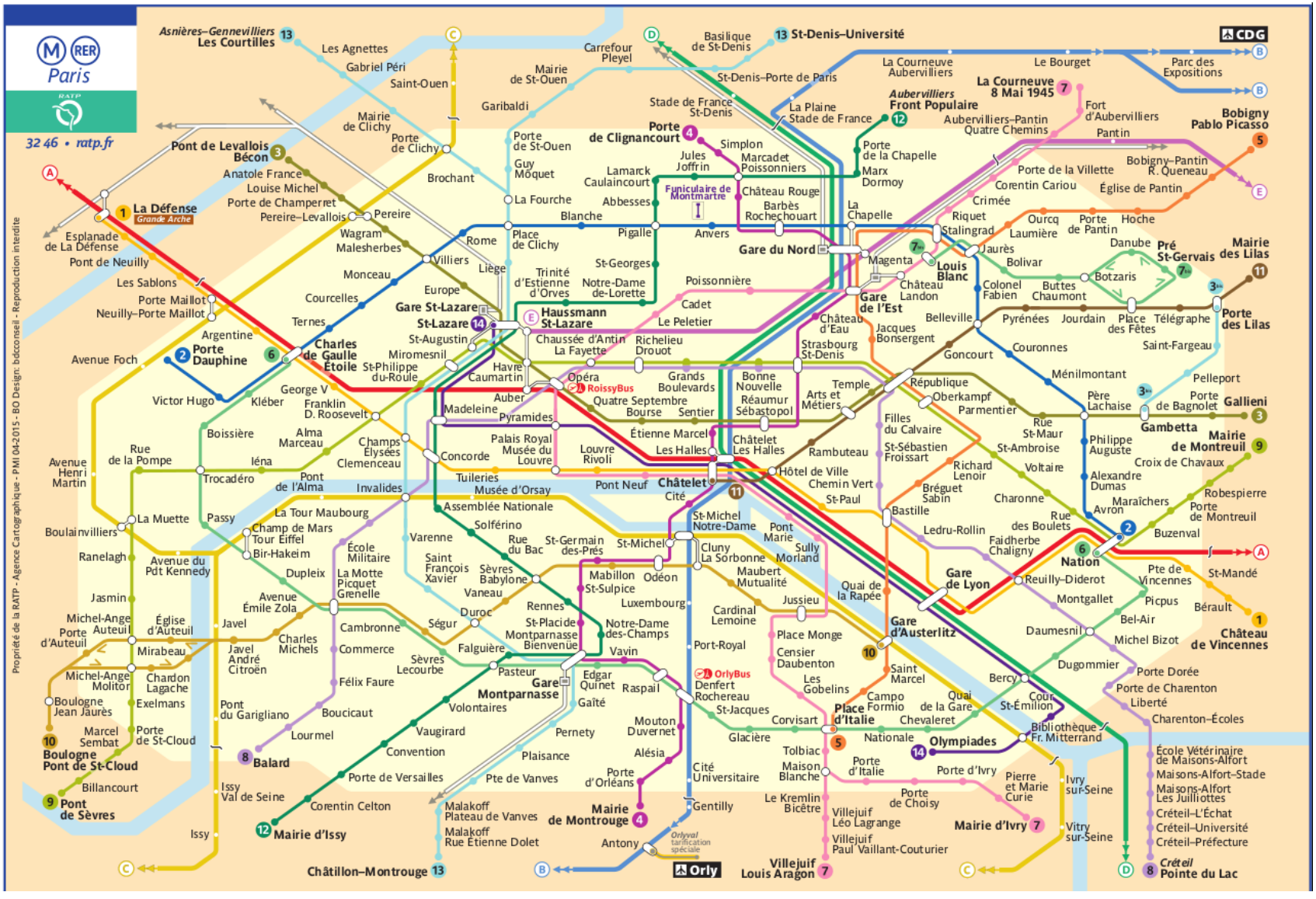
Objectives

- Aggregate openly available data
- Connect the data to create a single large graph
- Analyse the more global structure, make the data easily available for research and applications.

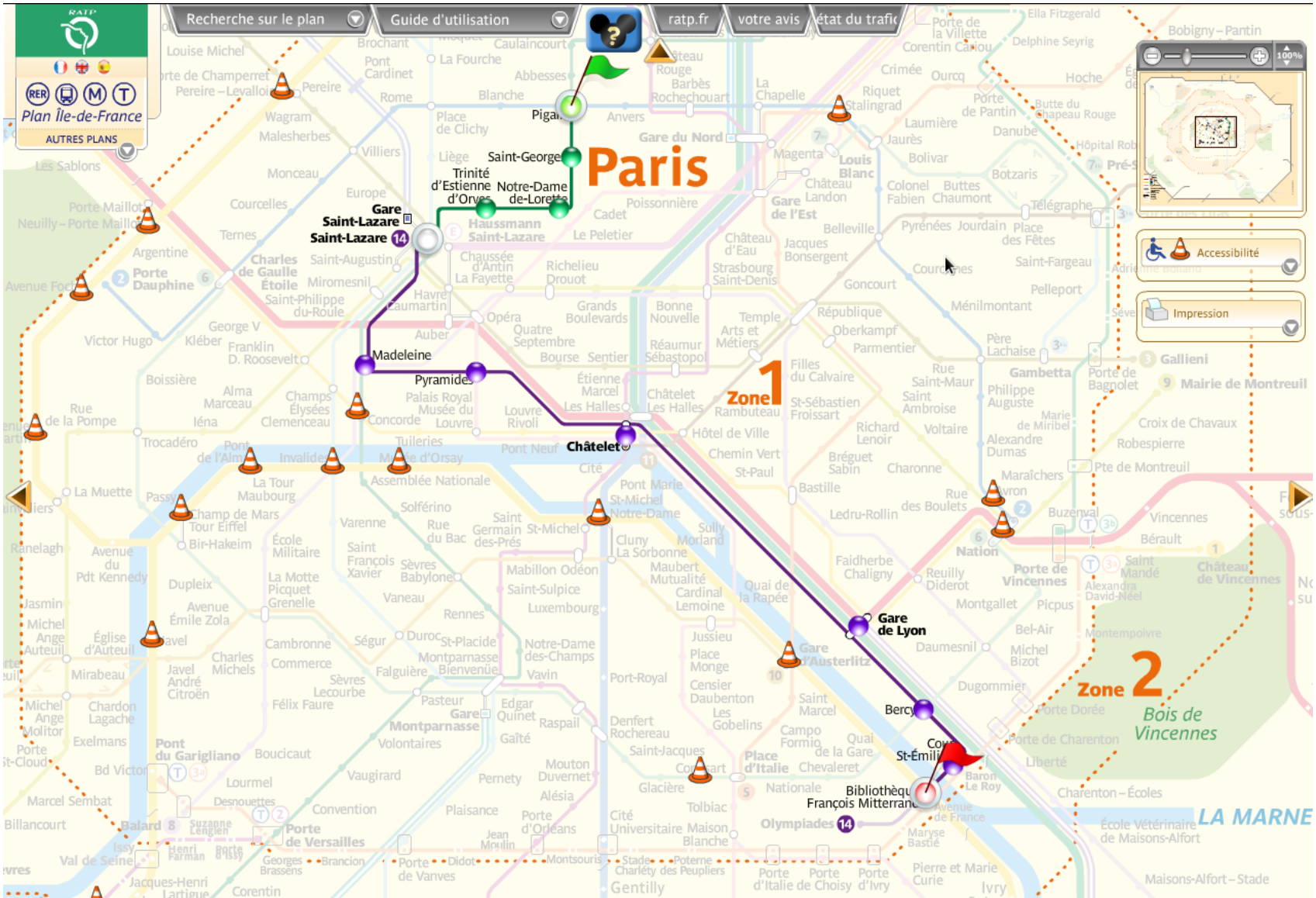
What is a graph

A graph is a theoretical object made of Vertices and Edges.
Edges can be directed.
In some cases edges and vertices can have properties (similar to some feature structure).
A large body of works in mathematics focus on graph analysis, which results in new tools readily available for other disciplines.

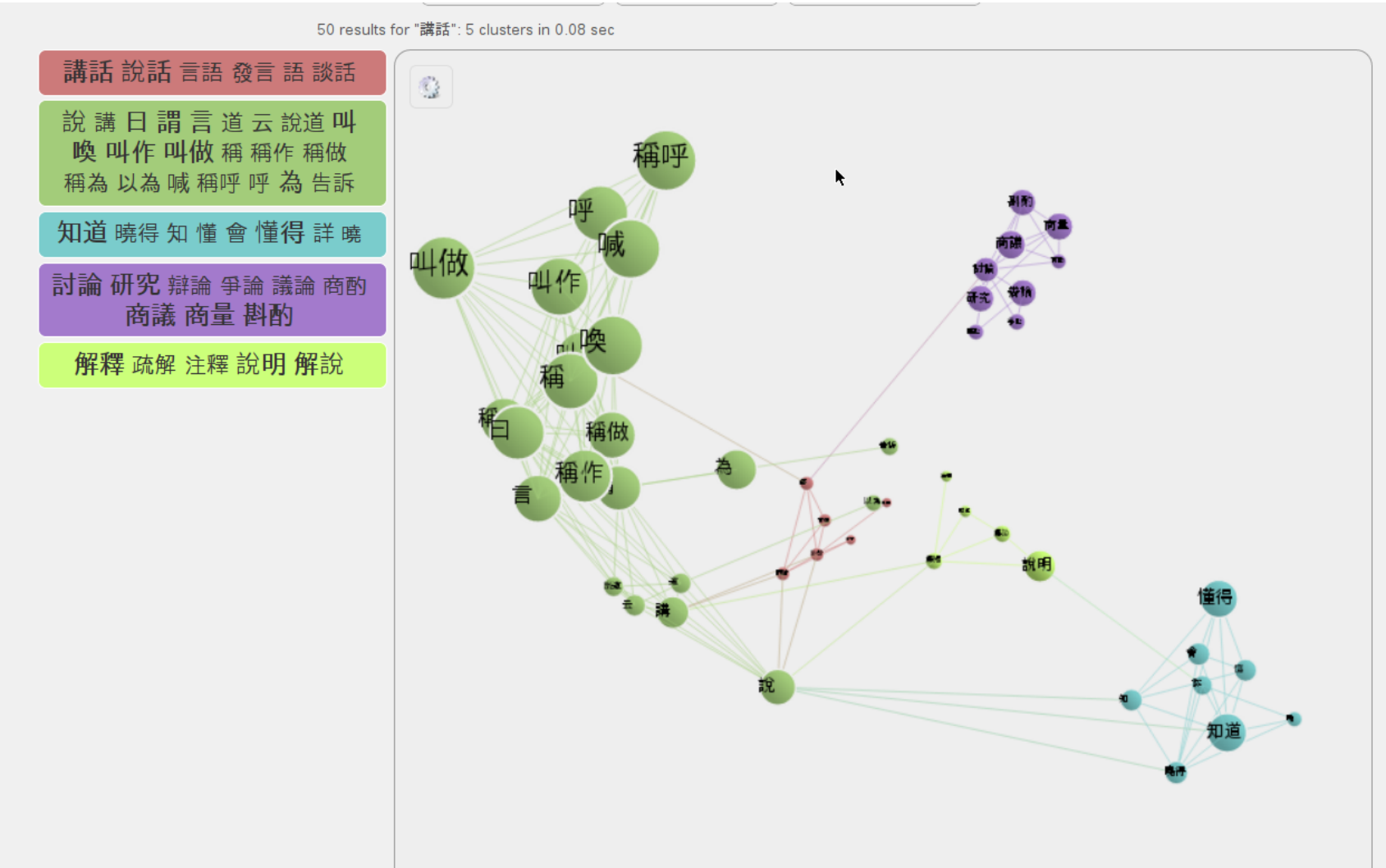
Example of a familiar graph



Algorithm on graphs: Shortest Path



Algorithm on graphs: Community Detection (Tmuse, Chudy et al. 2013)

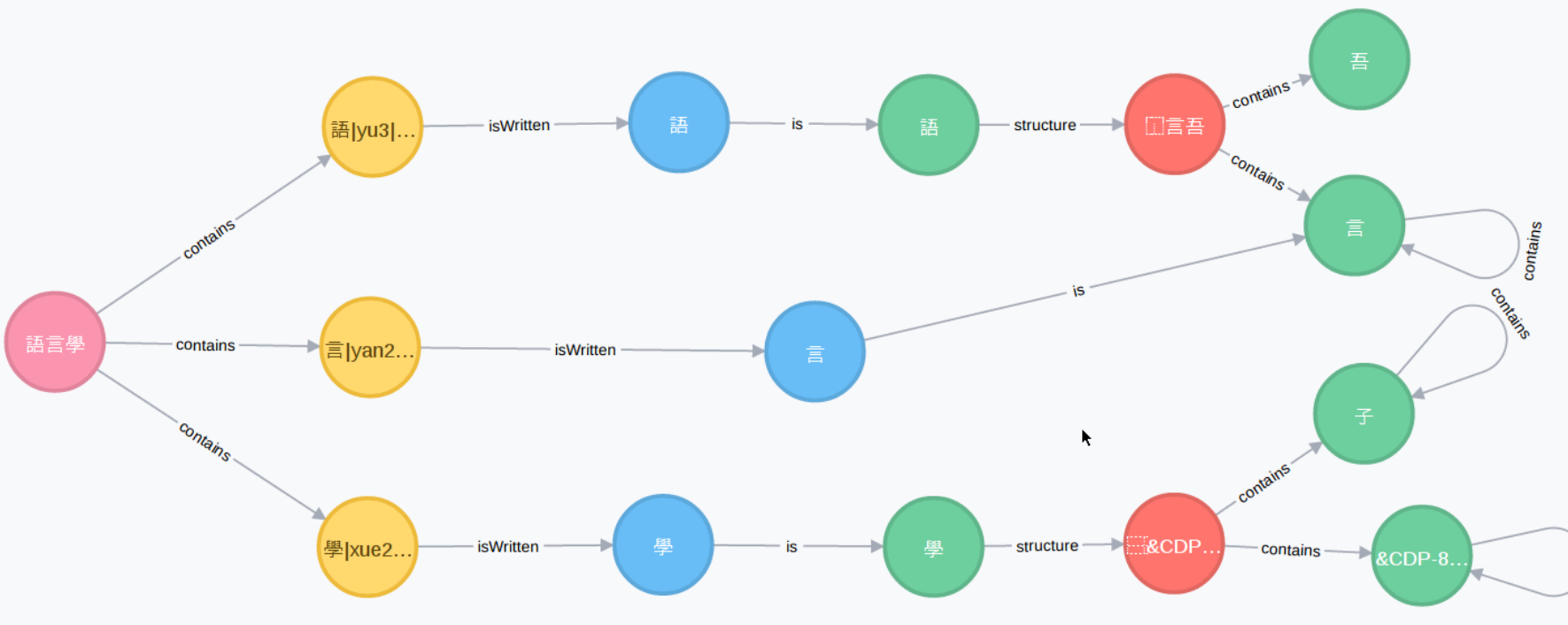


A Glance at our Graph

```
$ MATCH (e:Expression {sino:'語言學'}), (e)--> (l:Lecture), (l) --> (s:Sinogram), (s) --> (c:Component) WITH e,l
```

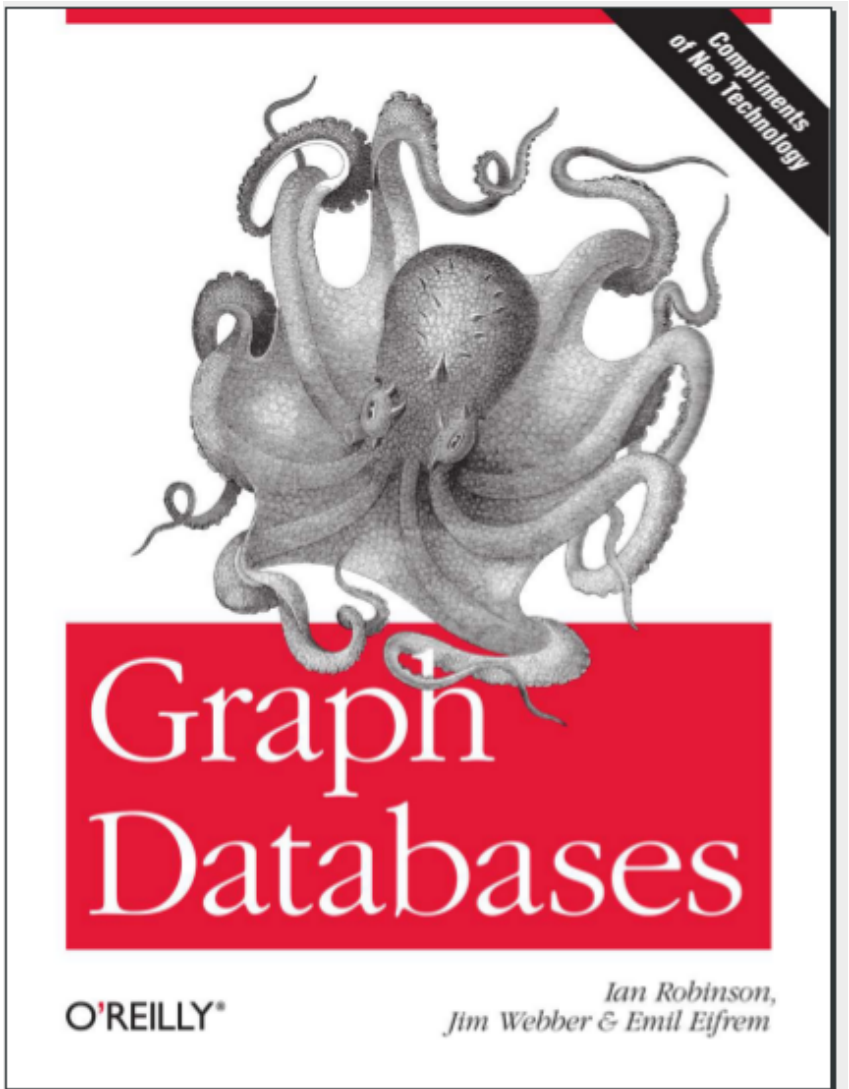
*(15) Component(6) Expression(1) IDS(2) Lecture(3) Sinogram(3)

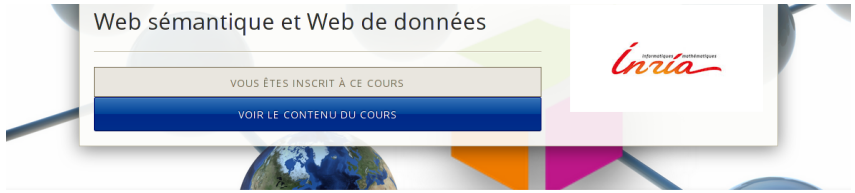
*(18) contains(10) is(3) isWritten(3) structure(2)



How to model your data as a graph

keywords: GraphDB, Semantic Web





présentation

A PROPOS DU COURS

Ce cours vous propose de vous former aux standards du Web de données et du Web sémantique. Il vous présentera les langages qui permettent :

- de représenter et de publier des données liées sur le Web (RDF) ;
- d'interroger et de sélectionner très précisément ces données à distance et au travers du Web (SPARQL) ;
- de représenter des vocabulaires et de raisonner et déduire de nouvelles données pour enrichir les descriptions publiées (RDFS, OWL, SKOS) ;
- et enfin, de tracer et de suivre l'histoire des données (VOiD, DCAT, PROV-O, etc.).

À QUI S'ADRESSE CE COURS ?

Ce cours s'adresse à des étudiants ou ingénieurs en informatique, notamment dans le domaine des systèmes d'information. Il peut être suivi en complément d'une formation classique aux technologies de base du Web.

PRÉ-REQUIS

MOOC de niveau Licence, à destination de personnes ayant des connaissances de base en informatique, notamment sur les outils et langages classiques du Web (navigateurs Web, HTML de base, etc.), et la

Inscriptions	**Fermé**
Début du Cours	**02 mars 2015**
Fin des cours	**17 avril 2015**
Effort estimé	**02:00 h/semaine**

Previous works on lexical networks

Some sources of inspiration:

- Studies on Synonyms Networks by Bruno Gaume
- Réseau Lexical du Français (from RELIEF Project, headed by Alain Polguère)
- BabelNet (ERC MultiJEDI headed by prof. Roberto Navigli)
- efforts towards Linguistic Linked Open Data

List of data sources used

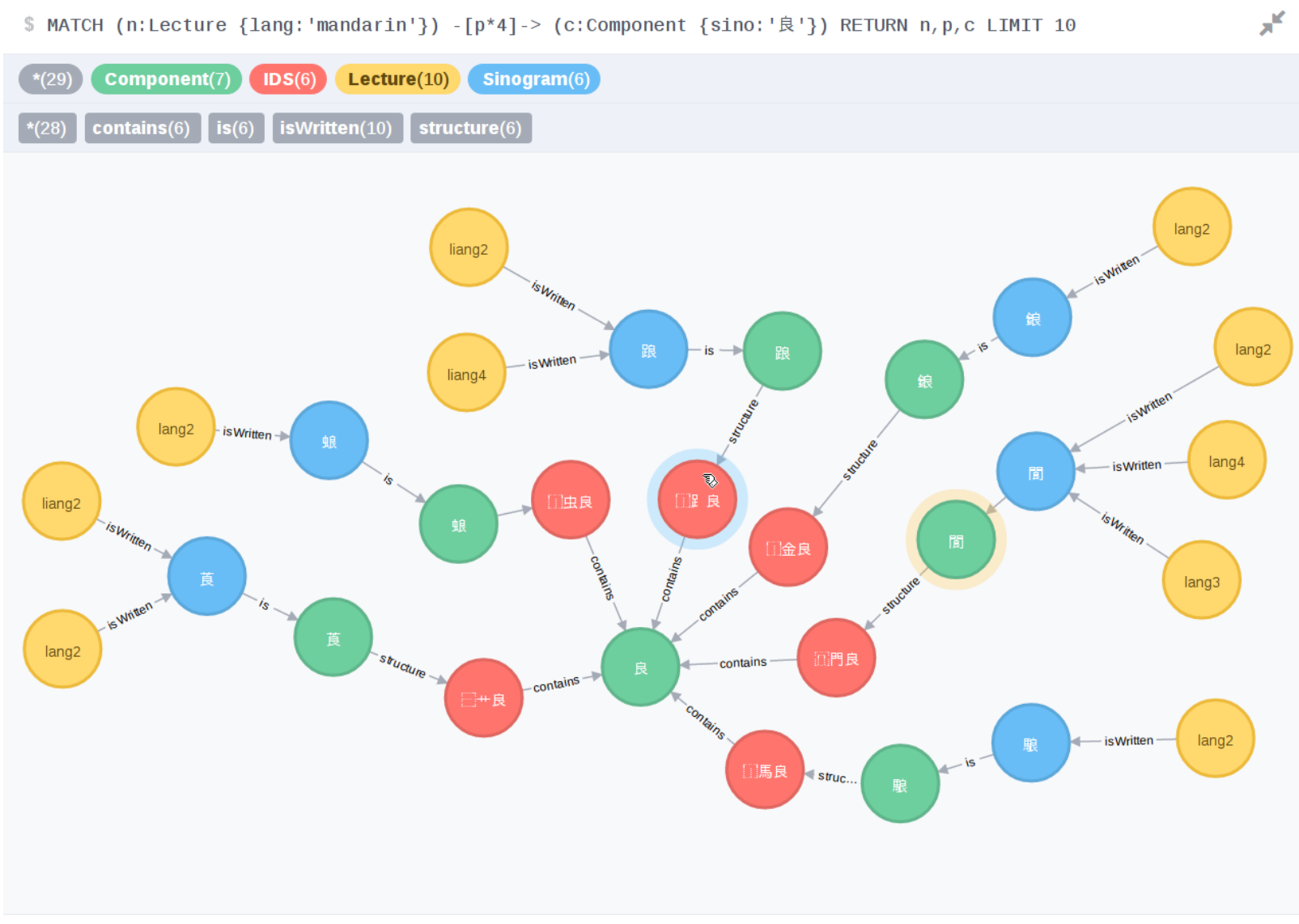
(or that we plan to use)

- “Chine Au Logis” (Marc & Magistry)
- 萌典 (Taiwan’s MOE and g0v.tw)
- 『臺灣閩南語常用詞辭典』 (Taiwan’s MOE)
- Various languages tests programs like HSK
- Chinese Wordnet (and some others to come)
- classics (說文解字、廣韻)
- IDS
- Baxter & Sagart reconstructions (2011,2014)
- Tmuse synonyms graphs
- Wiktionaries (DBnary)
- Tatoeba.org
- CEDict (and German and French descents)
- JEDict
- grammatical information extracted from open corpora
- Unicode metada (unihan

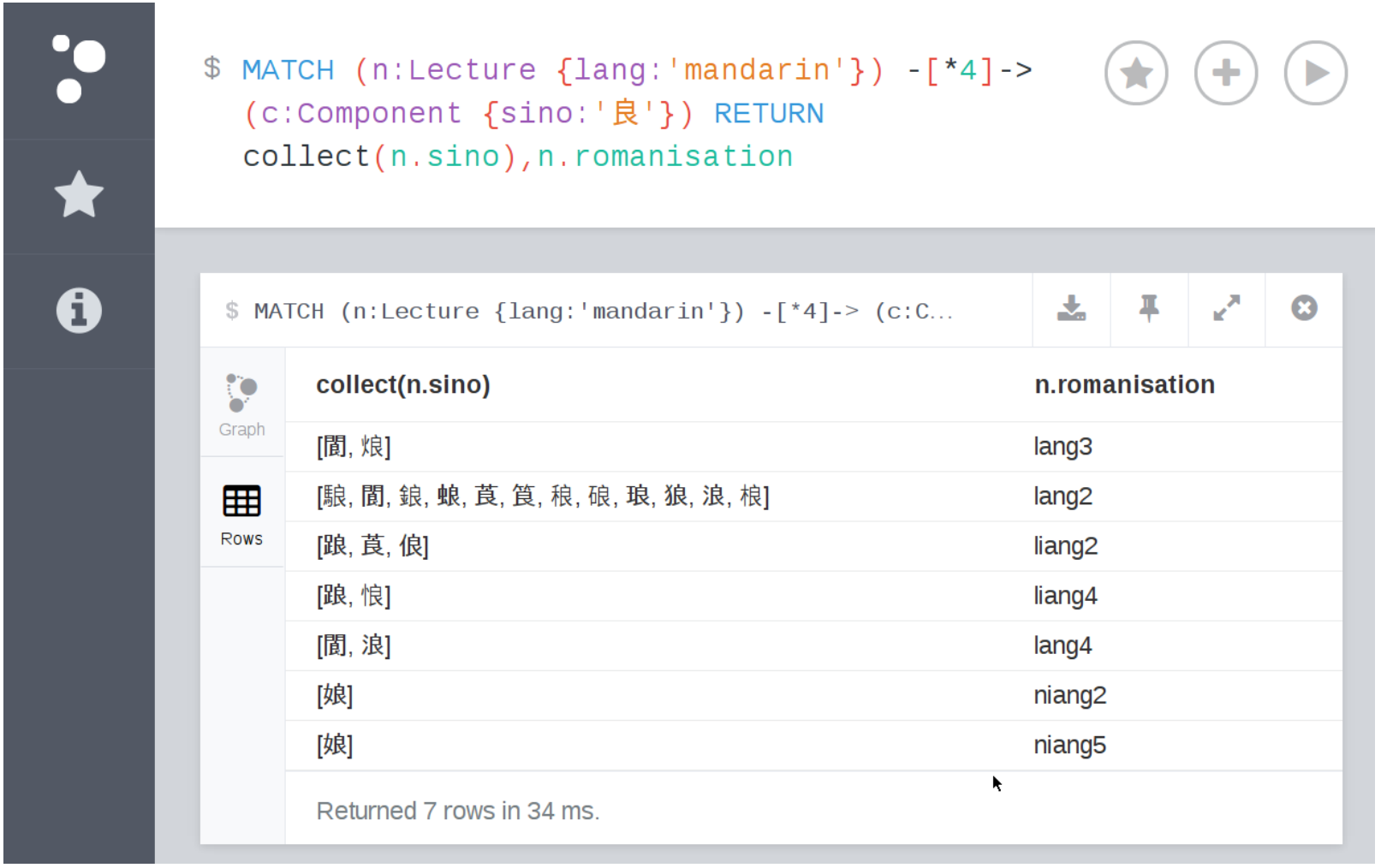
Sinograms

- strokes: numbers, orders
- structures:
- components
- grammotological variants
- synchronic graphic variants
- readings
- fanqie
- rhymes
- transcriptions
- phonological features
- grammatical information extracted from open corpora
- Unicode metada (unihan database)

Graphical decomposition



Graphical decomposition



Graphical decomposition

Lexical Units

- Language dependent part
- grammatical information
- Any relation you could find in a Wordnet or other Lexical Network
- Translation links

Licence and Distribution

Some methodological and legal aspects of our work:

- We aggregate only Open Data
- We don't create new data, except for
  - adding relations between existing datasets
  - extracting information from corpora automatically
- we will distribute the code to aggregate the data
- the licence of the output thus depends on the input

Export as a Semantic Web graph

(work in progress)

TODO:

- some cleaning (homogeneisation)
- define the URIs for our objects
- documentation about the terms used in the graph
- publish it !
- get feedback (hopefully)

Stratification of lexical borrowings

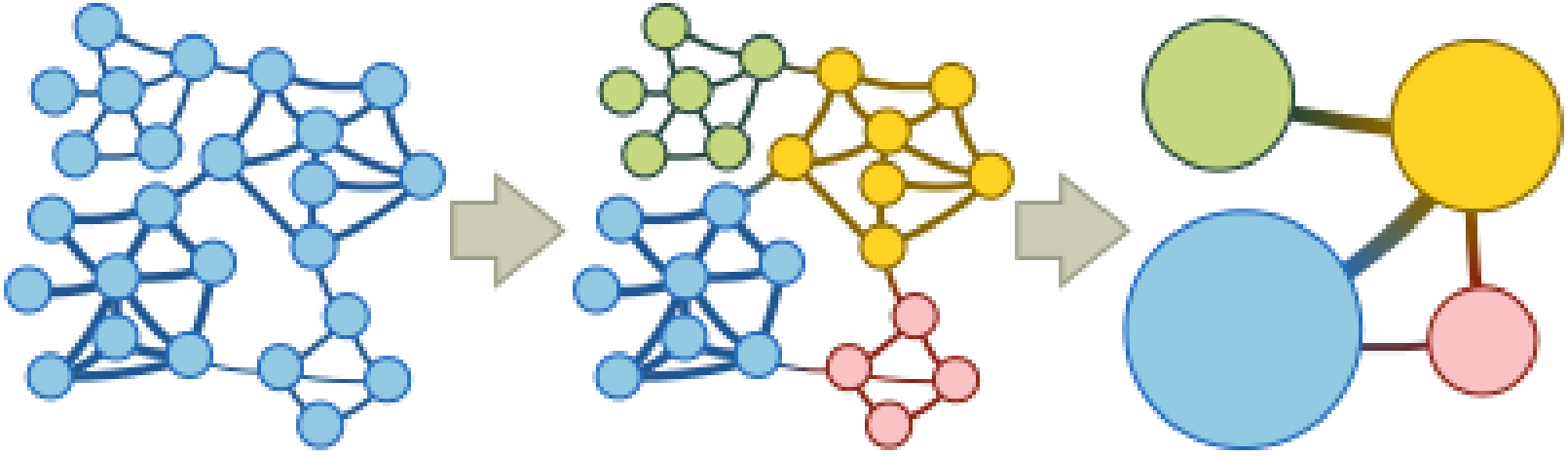
(Magistry, 2015)

Using this dataset, we were able to

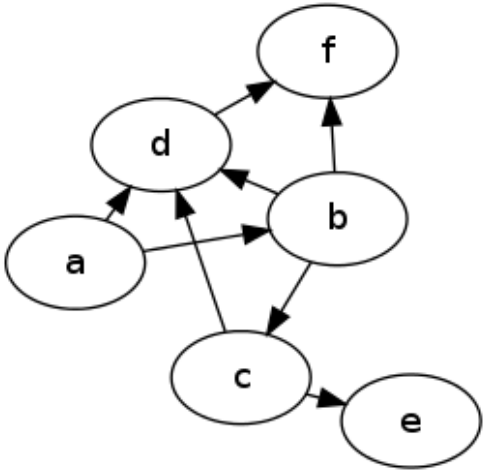
- Implement the Extended Principle of Coherence (Sagart and Xu, 2001) on a large scale
- Clusterize the readings of sinograms that are expected to be coocurring in time
- Define an order on the clusters based on the 文/白讀音 traditional analysis

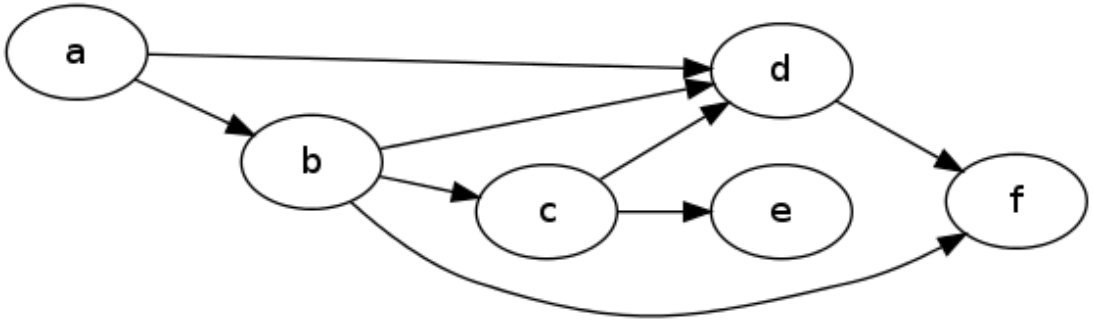
Stratification of lexical borrowings

clusterisation



layout





Stratification of lexical borrowings

to be continued with

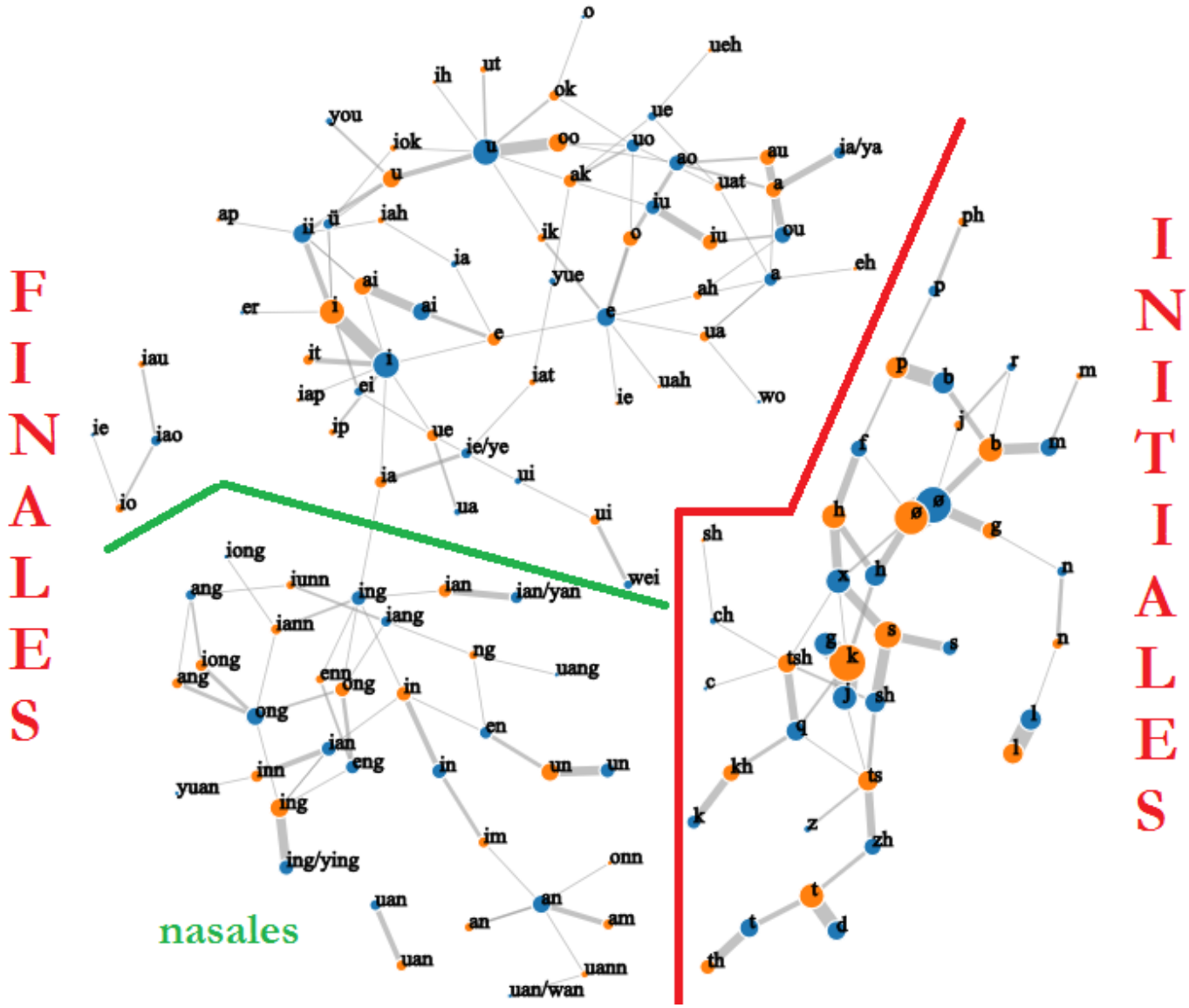
- with comparison with the result with the phonology
- extension of the work by using phonetic component rather than whole sinograms

Study on Phonetic components in synchrony

In an effort to ease language learning, we

- extract all readings related to a component
- automatically classify the components into phonetic or non-phonetic (based on the entropy of the probability distribution of the readings)
- we can do this multi- or crosslingualy

Sinogramic Transposition for Mutual Understanding



Bibliography I