UQÀM

# The Dictionary Game

## Toward a Characterization of Lexical Primitives Using Graph Theory and Relational Concept Analysis

Presented by
Mickaël Wajnberg

Joint work with :
Poulin Jean-Marie
Blondin-Massé Alexandre
Valtchev Petko

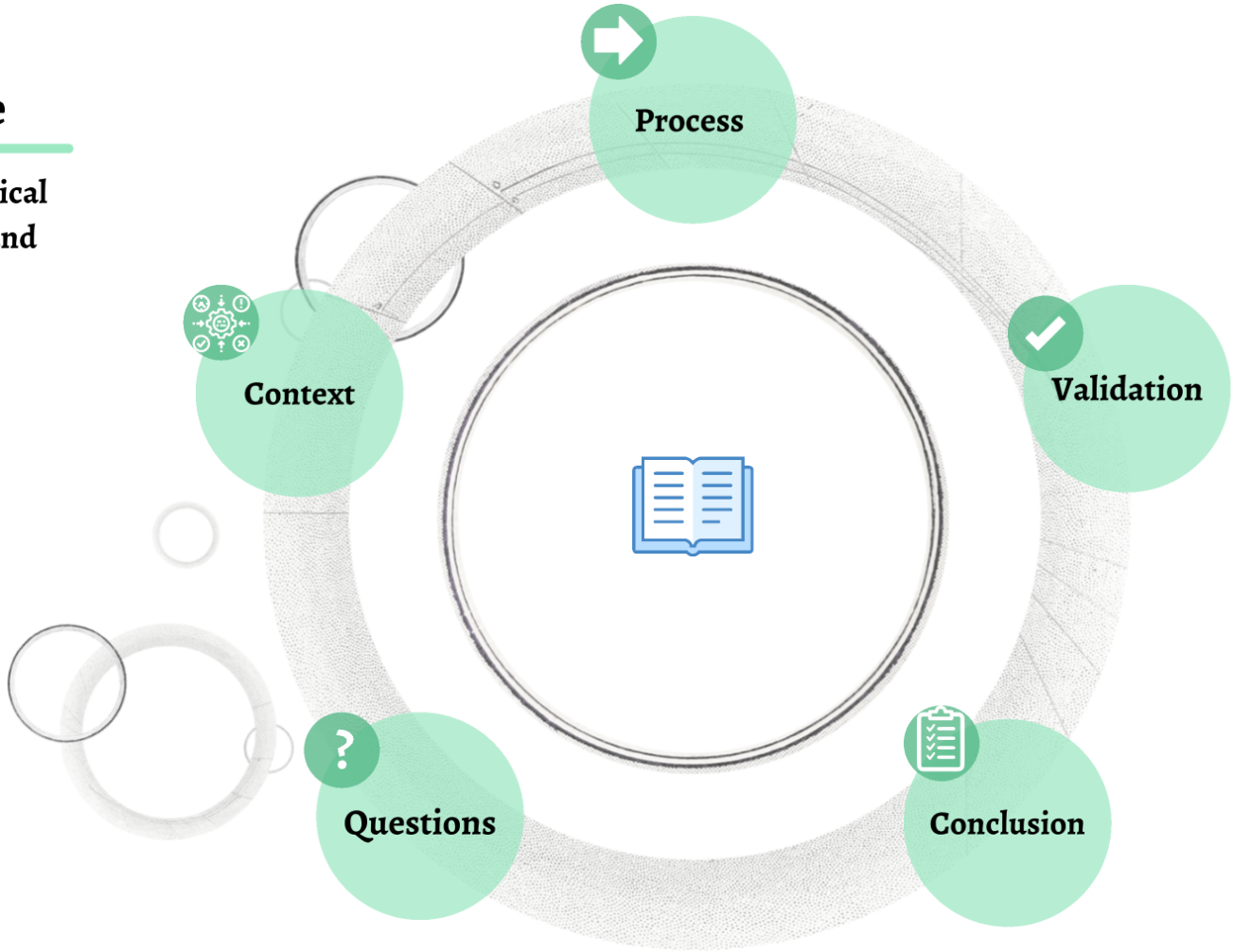
contact : wajnberg.mickael@courrier.uqam.ca
link for interactive presentation : http://bit.ly/COG20-WAJ-prez

# About the presenter :

Mickael Wajnberg is a student, currently enrolled in a PhD at University of Quebec at Montreal (Québec, Canada) and at Université de Lorraine (France), he currently works on Relational Concept Analysis and knowledge extraction. He did a Math and Physics Prepa before he got an Engineering Degree (M. Sc equivalent) at Telecom Nancy (France) and a M. Sc at University of Quebec at Chicoutimi (Québec, Canada) in Computer Science, he specialized in algorithms and theory for computer science.

# The Dictionary Game

**Toward a Characterization of Lexical Primitives Using Graph Theory and Relational Concept Analysis**



# Symbol Grounding Problem


Learning a new language


Using only a dictionary


Linking with known words


Expand to full dictionary

Is there an optimal set
of words to link first ?

# Lexical Primitives

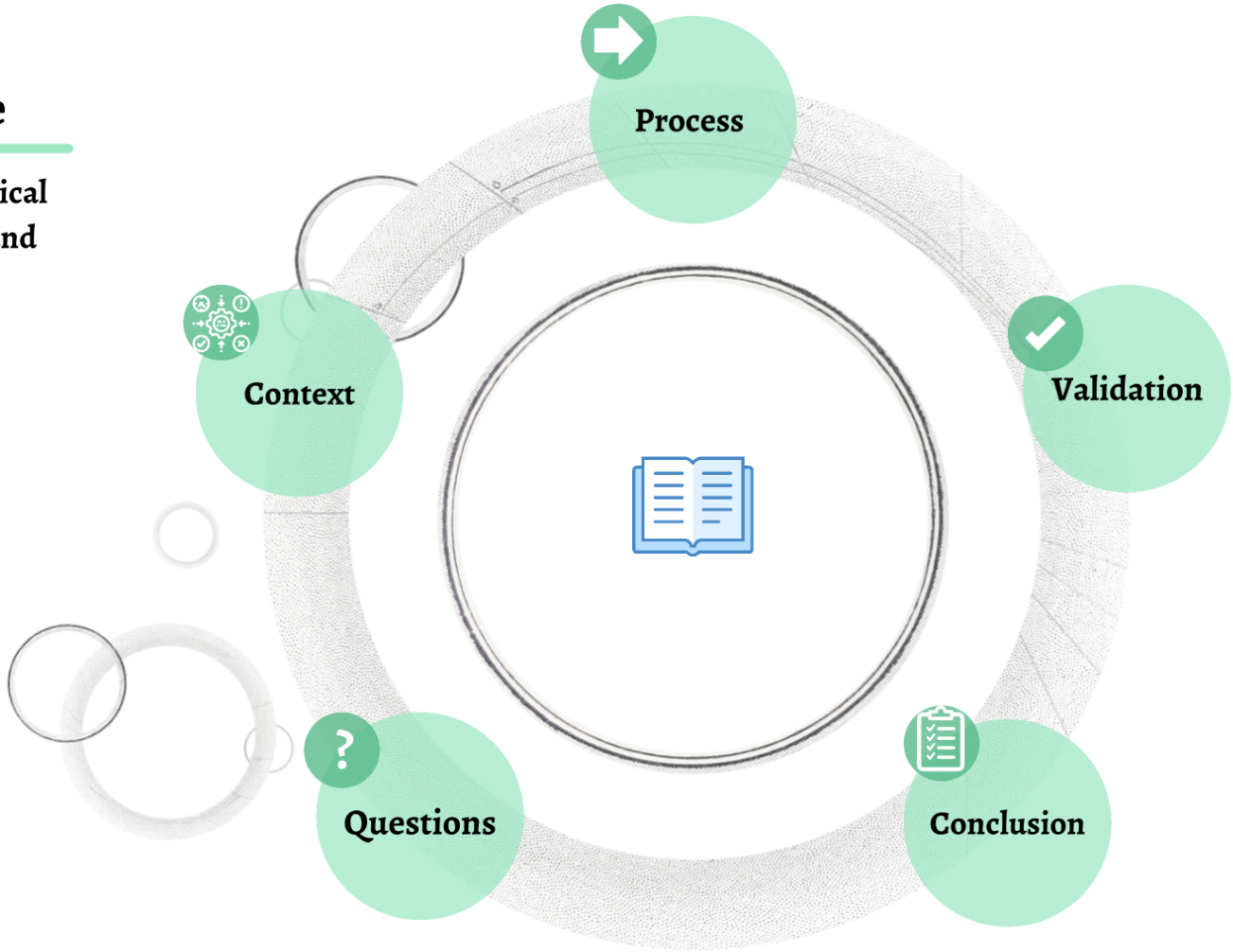
Set of words expanding to dictionary by relation of definition

Minimal Grounding Set (MGS)

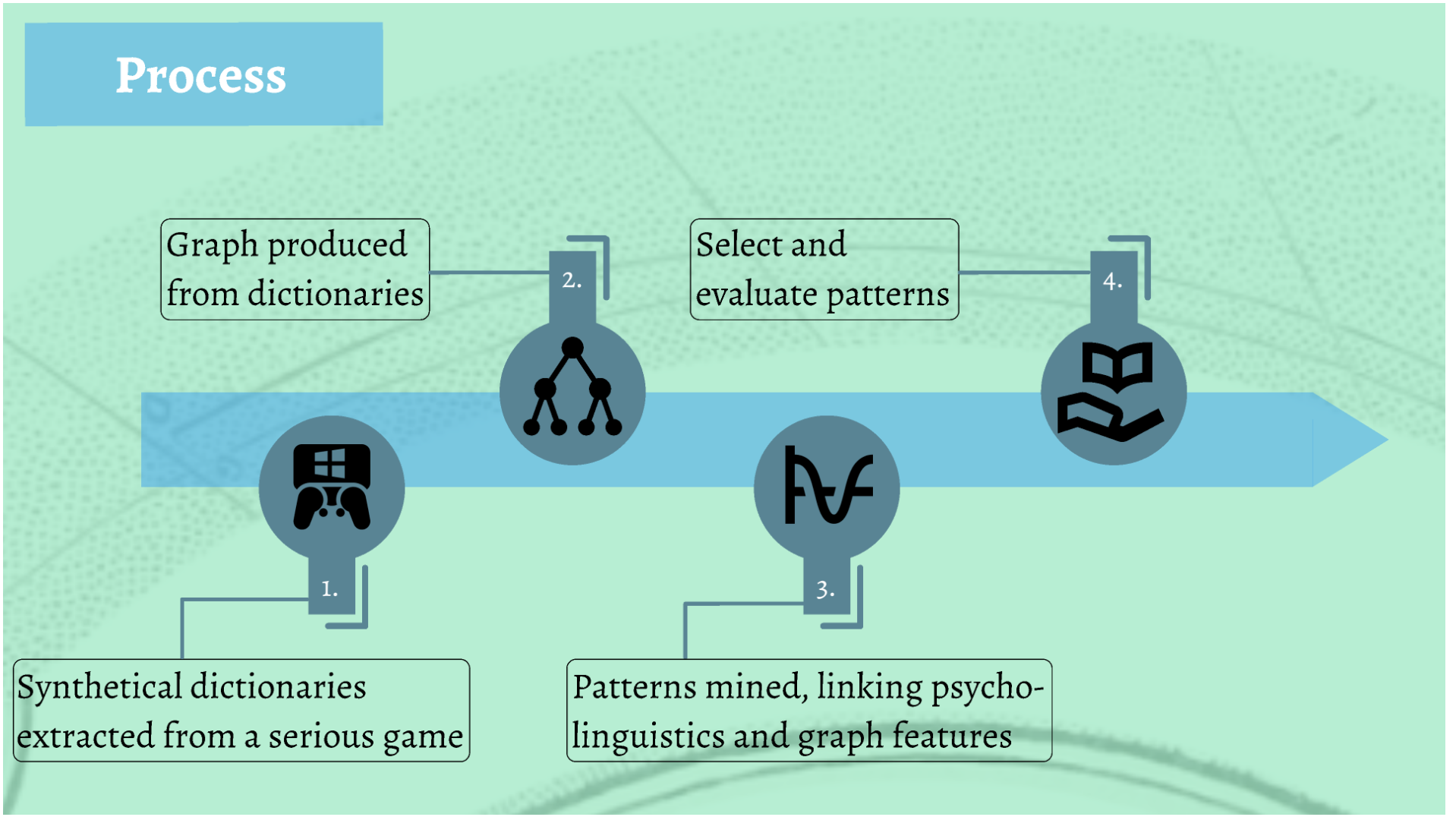
MGS words share psycholinguistics characteristics

# The Dictionary Game

**Toward a Characterization of Lexical Primitives Using Graph Theory and Relational Concept Analysis**



# Process


Graph produced from dictionaries
2.
Select and evaluate patterns
4.
1.
3.
Synthetical dictionaries extracted from a serious game
Patterns mined, linking psycho-linguistics and graph features

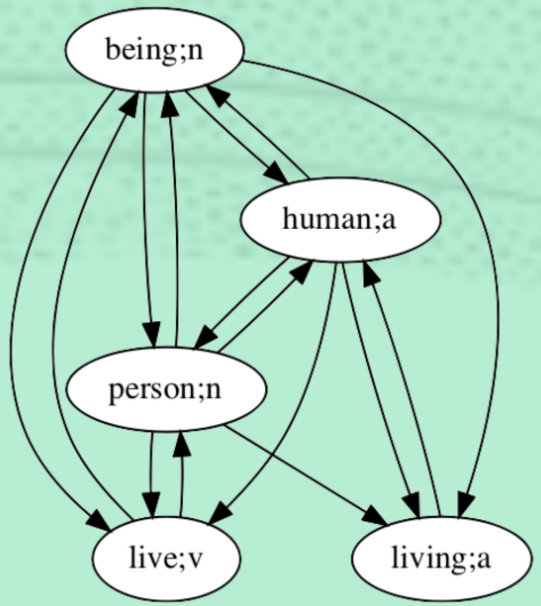

**4 root words available :**

- **clock**
- **horse**
- **person**
- **thing**

# Graph Representation

Given a set of definitions,
a graph is given as an ordered pair (V,E) where :

- V : the set of words
- E : (x,y) in E, if and only if x occurs in the definition of y



# Association Rules

Information format

**Template :**

**A ⟶ B**

Example :

A = {Large Diameter, >70% abstract words}

B = {Few Strongly Connected Components}

**Support : # objects with A & B**

**Confidence : % of objects having B among these verifying A**

**Why ?**

**Understandability**

**Quantifiable**

**No label required**

**Relational data compatible**

# Association Rules

Extraction

## Formal Concept Analysis :

- **Groups objects sharing common features**
  Ex : Concrete and frequent words

- **Organize the groups in a lattice hierarchy**
  Ex : Subset of concrete words

- **Representation base can be extracted**
  Ex : Frequent ⟶ Concrete

# Association Rules

Extraction

## Relational Concept Analysis :

- **Extends FCA to relational data**
  Ex : Dictionaries(T), Words(T), Contain(R)

- **Propositionalize groups into attributes**
  Ex : contains "at least one of" Member(concrete words)

- **Multiple propositionalization operators available**
  Ex : "at least one of", "all", "at least p% of"...

# Validation

Application to the dictionary game data

**~ 100 Dictionaries**

6 structural features

**~ 2500 Words**

5 psycholinguistic metrics

**Contains Relation**

~125 word/dictionary

## Results

# Results

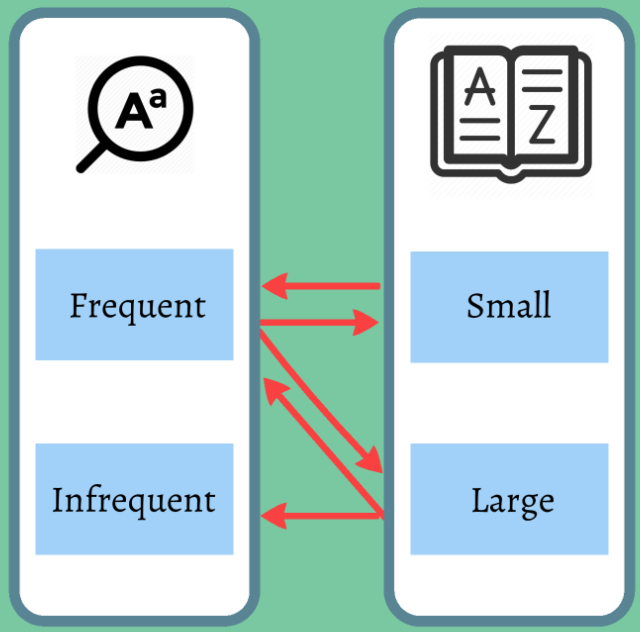
## Association Rules interpretation

### Observations

- Frequent words are used in every dictionaries
- Small dictionaries are limited to these words
- Large dictionaries also include infrequent

### Conclusions

- For each root words there is an ideal set of words to build the dictionary
- When faced to unknown/hard definitions, two phenomena can occur :
  - A synonym is used
  - Unprecise definitions generate noise



# Validation

Application to the dictionary game data

**~ 100 Dictionaries**

6 structural features

**~ 2500 Words**

5 psycholinguistic metrics

**Contains Relation**

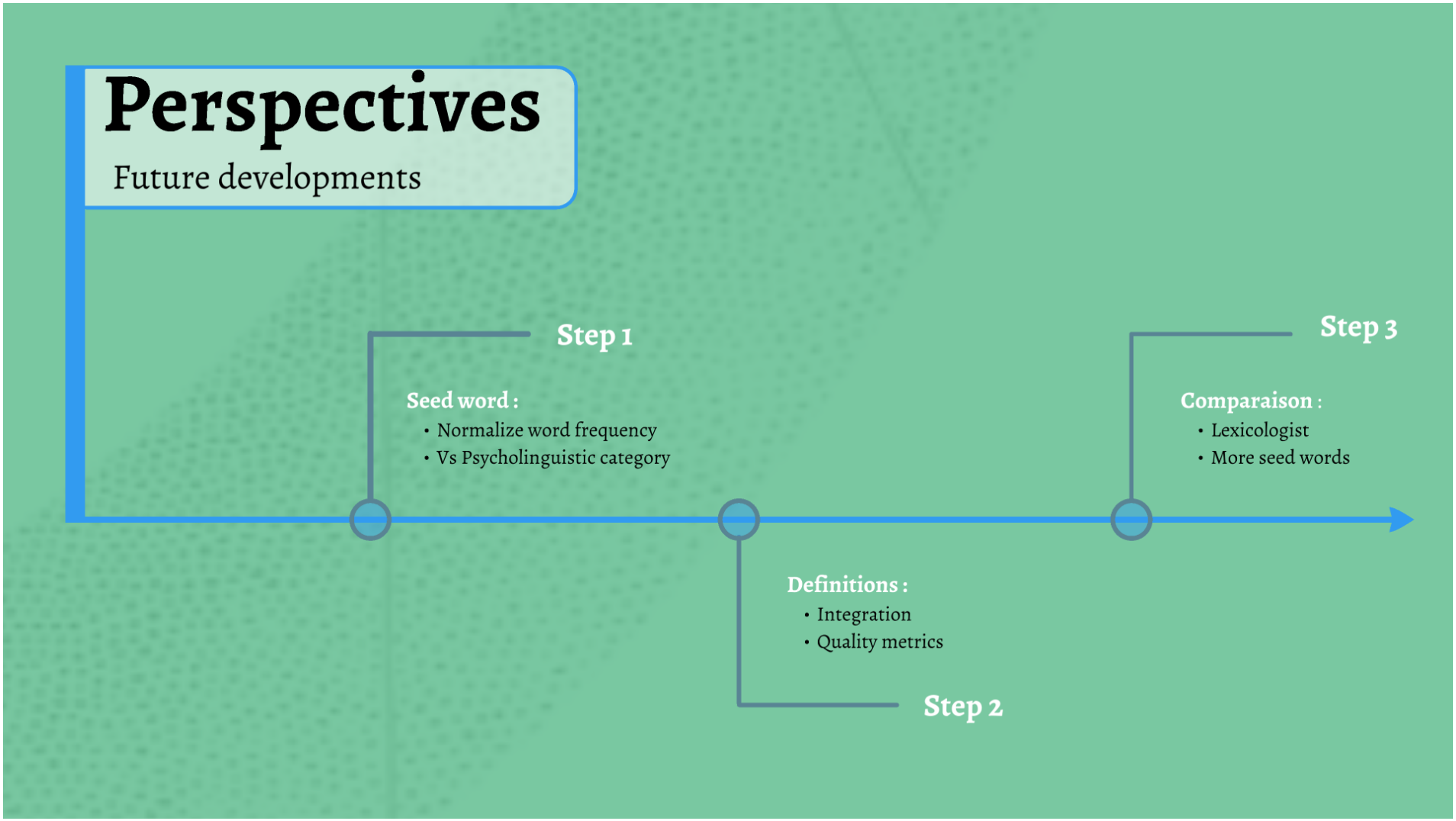
~125 word/dictionary

## Results

# Conclusion

- ✓ Ideal set per root
- ✓ Higher density
- ✓ Players strategies
- ✗ No association with age of acquisition
- ✓ Minimize dict. size
- ✓ Especially in the largest SCC
- ✗ No association with concreteness
- ✗ Number of words is a limited metric

## Perspectives


Perspectives
Future developments
Step 1
Seed word :
• Normalize word frequency
• Vs Psycholinguistic category
Step 3
Comparaison :
• Lexicologist
• More seed words
Definitions :
• Integration
• Quality metrics
Step 2