SELEXINI

SEmantic **LE**Xicon **IN**duction for **I**nterpretability and diversity in text processing

Induction de lexiques sémantiques pour l'interprétabilité et la diversité en traitement de textes

Projet financé par l'Agence Nationale de la Recherche
2022-2025

Program

Thursday July 07 (Luminy campus)

- 14:30-15:30 Invited talk by Denis Paperno
- 15:30-16:00 Overview of the project
- 16:00-16:30 Tour de table
- 16:30-17:00 coffee break
- 17:00-18:30 Working group on WP1
- 18:30-21:00 ~~Hike+picnic+swim in Sugiton~~ Picnic + swim on the Prado beach ☀

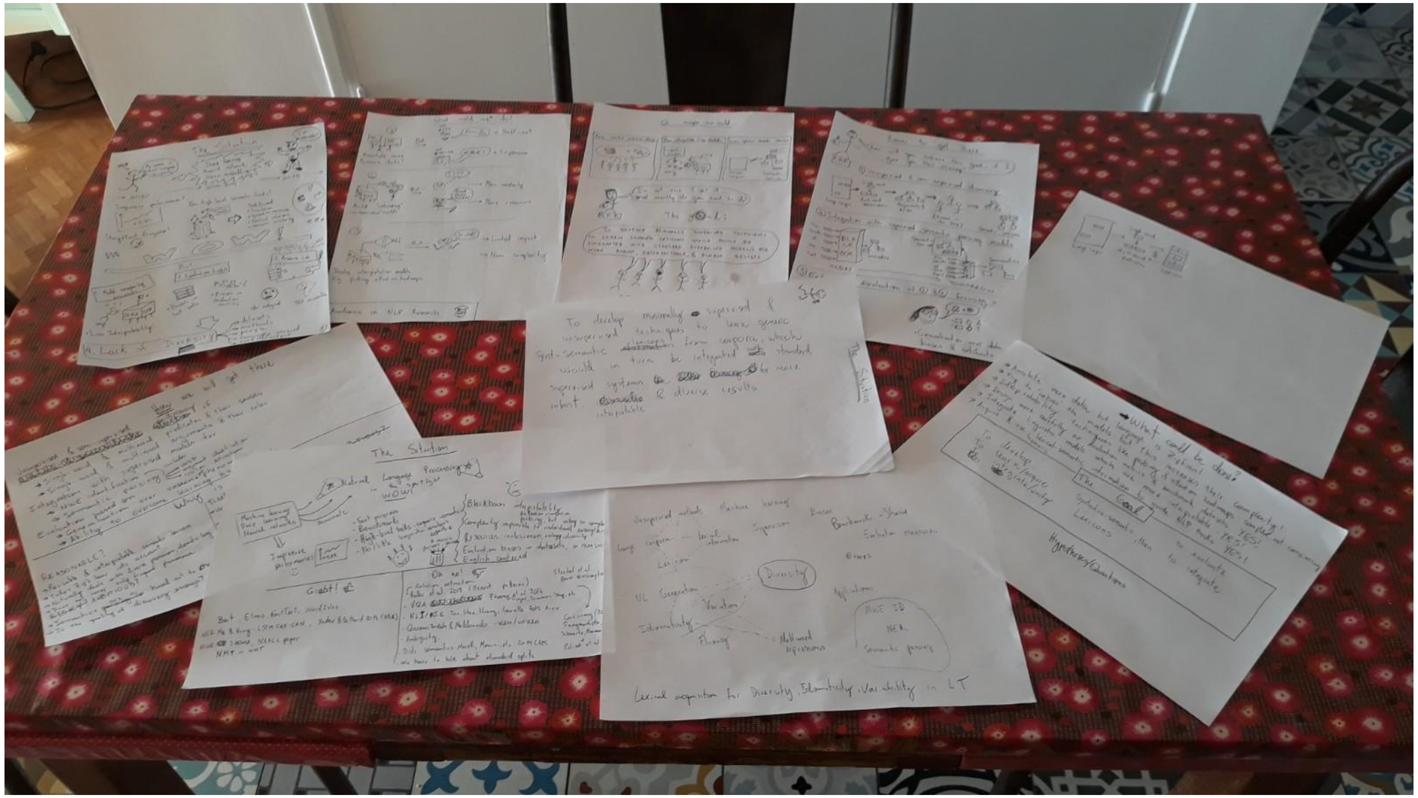
Friday July 08 (St Charles campus)

- 9:00-10:00 Working group on WP1
- 10:00-11:00 Working group on WP2
- 11:00-11:30 coffee break
- 11:30-12:30 Working group on WP5
- 12:30-14:00 Lunch

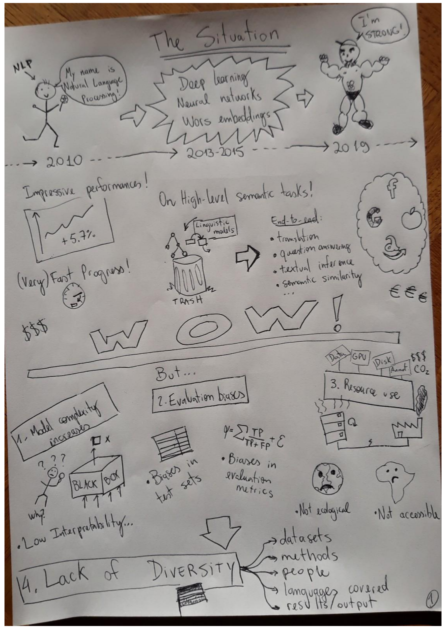
Where do we come from

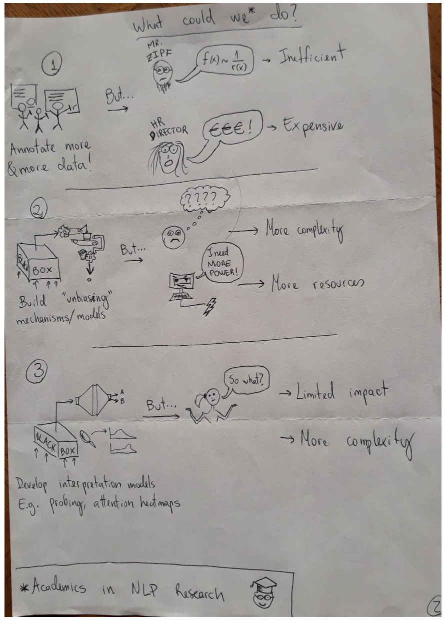
- PARSEME COST Action
- PARSEME-FR project
  - Parsing and **multiword expressions** in French

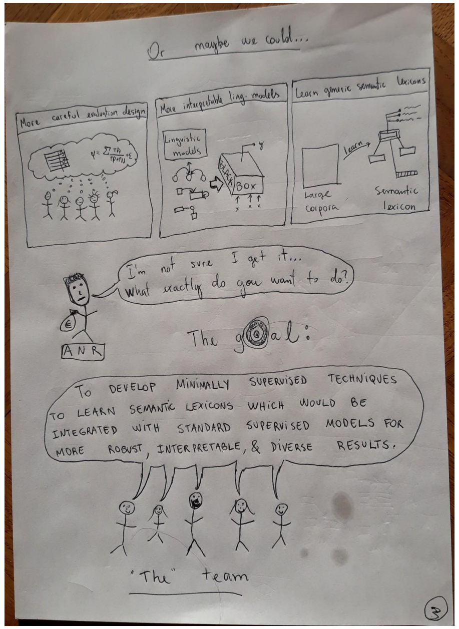
The birth of SELEXINI

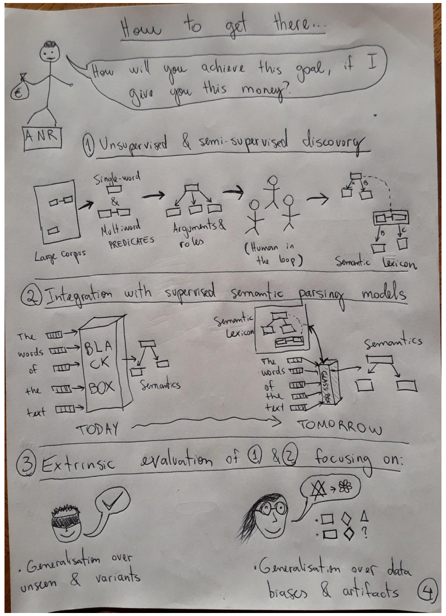

The Solution
The Goals
To develop minimally supervised & unsupervised techniques to learn generic synt-semantic lexicons from corpora, which would in turn be integrated with standard supervised systems for more robust, interpretable & diverse results
The Situation
Natural Language Processing
WOW!
Diversity
MWE ID
NER
Semantic parsing
Lexical acquisition for Diversity, Idiomaticity, Variability in LT
What could be done?
The Goal
Hypotheses/Questions

The birth of SELEXINI






Or maybe we could...
More careful evaluation design
$\varphi = \sum \frac{TP}{TP+TN} + \mathcal{E}$
More interpretable ling. models
Linguistic models
y
BLACK BOX
x x x
Learn generic semantic lexicons
learn
Large corpora
Semantic lexicon
I'm not sure I get it... What exactly do you want to do?
€
ANR
The goal:
TO DEVELOP MINIMALLY SUPERVISED TECHNIQUES TO LEARN SEMANTIC LEXICONS WHICH WOULD BE INTEGRATED WITH STANDARD SUPERVISED MODELS FOR MORE ROBUST, INTERPRETABLE, & DIVERSE RESULTS.
"The" team
3


How to get there...
€
How will you achieve this goal, if I give you this money?
ANR
1 Unsupervised & semi-supervised discovery
Large corpus
Single-word & Multiword PREDICATES
Arguments & roles
(Human in the loop)
A B
B C
Semantic Lexicon
2 Integration with supervised semantic parsing models
The words of the text
BLACK BOX
Semantics
Semantic Lexicon
The words of the text
GLASS BOX
Semantics
TODAY
TOMORROW
3 Extrinsic evaluation of 1 & 2 focusing on:
• Generalisation over unseen & variants
?
• Generalisation over data biases & artifacts
4

NLP today: between enthusiasm…

- Continuous representations highly adapted to neural models
- Transfer learning by fine-tuning large language models pre-trained on raw text using self-supervision
- Significant and regular performance improvements on all tasks
- End-to-end approaches possible, bypassing the need for traditional linguistic analysis

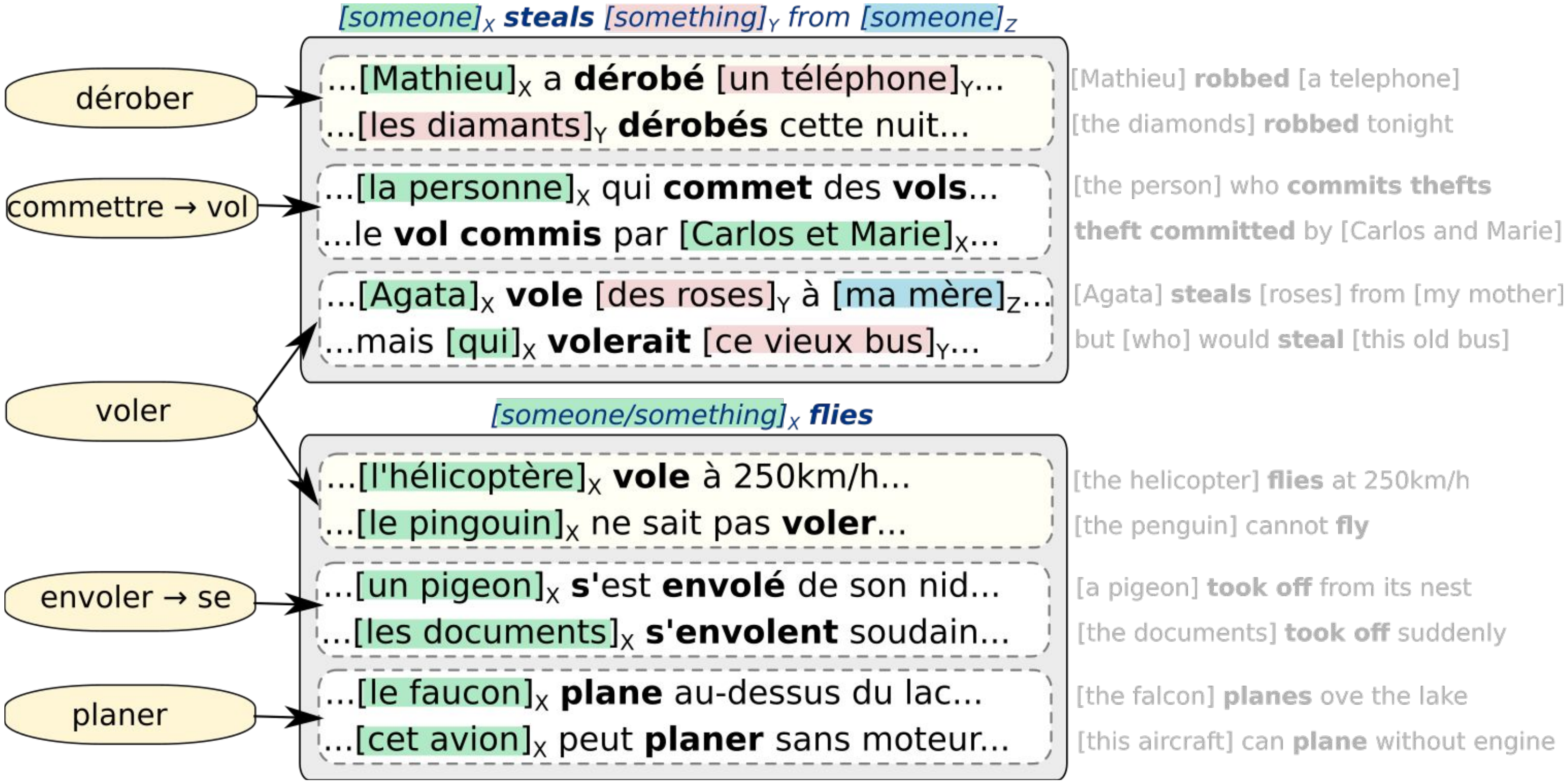
NLP today: ...and limitations

- **Model opacity**: millions (or billions) of real-numbered parameters
- **Lack of diversity**:
  - Repeated evaluation on potentially **biased** benchmarks
  - **Frequent** phenomena are favoured over rarer ones
  - Increase **performance** in spite of robustness
- **Implicit compositionality representations:**
  - Regular composition, e.g. argumental structure
  - Irregular composition, e.g. multiword expressions, idioms

The underlying hypotheses

- The notions of **lexicon** and **lexical units** are cognitively important
- Lexical-semantic notions of **senses** and **frames** provide useful generalisations
- **Explicitly** modelling lexical units brings **interpretability** to neural model's outputs

The (initial) lexicon model



*[someone]$_X$ **steals** [something]$_Y$ from [someone]$_Z$*

dérober

...[Mathieu]$_X$ a **dérobé** [un téléphone]$_Y$... — [Mathieu] **robbed** [a telephone]

...[les diamants]$_Y$ **dérobés** cette nuit... — [the diamonds] **robbed** tonight

commettre → vol

...[la personne]$_X$ qui **commet** des **vols**... — [the person] who **commits thefts**

...le **vol commis** par [Carlos et Marie]$_X$... — **theft committed** by [Carlos and Marie]

voler

...[Agata]$_X$ **vole** [des roses]$_Y$ à [ma mère]$_Z$... — [Agata] **steals** [roses] from [my mother]

...mais [qui]$_X$ **volerait** [ce vieux bus]$_Y$... — but [who] would **steal** [this old bus]

*[someone/something]$_X$ **flies***

...[l'hélicoptère]$_X$ **vole** à 250km/h... — [the helicopter] **flies** at 250km/h

...[le pingouin]$_X$ ne sait pas **voler**... — [the penguin] cannot **fly**

envoler → se

...[un pigeon]$_X$ **s'**est **envolé** de son nid... — [a pigeon] **took off** from its nest

...[les documents]$_X$ **s'envolent** soudain... — [the documents] **took off** suddenly

planer

...[le faucon]$_X$ **plane** au-dessus du lac... — [the falcon] **planes** ove the lake

...[cet avion]$_X$ peut **planer** sans moteur... — [this aircraft] can **plane** without engine

SELEXINI's ambitions

1. Develop techniques to **induce** semantic lexicons automatically
   - From raw corpora
   - Using semi-supervised clustering
     - Seeds = lexical units and example sentences from Wiktionary.fr
2. … and **use these lexicons** within neural NLP systems
   - MWE identification
   - Machine reading comprehension

SELEXINI's ambitions

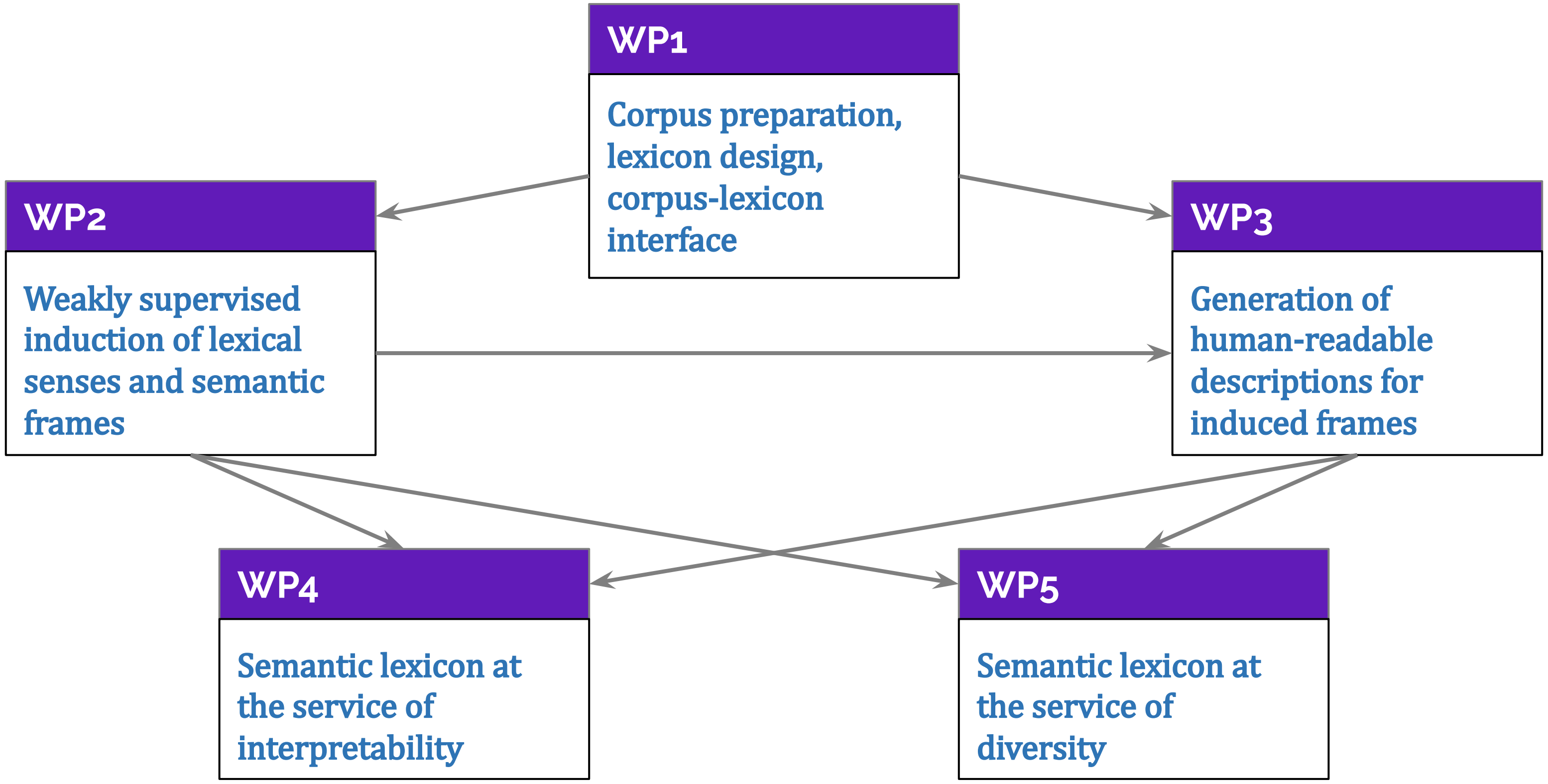
1. Develop techniques to **induce** semantic lexicons automatically
   - From raw corpora
   - Using semi-supervised clustering
     - Seeds = lexical units and example sentences from Wiktionary.fr
2. … and **use these lexicons** within neural NLP systems
   - MWE identification
   - Machine reading comprehension

- Lexical unit embeddings provide intermediate representations:
  - Static word embeddings: one vector per ambiguous lexical unit such as *voler*
  - Contextual word embeddings: a different vector for each occurrence of word *voler*

Why lexicon induction?

- Interpretability "by construction"
    - Hybrid continuous-symbolic model
- Large semantically annotated corpus as by-product
    - Lacking for many languages, including French
- High coverage with respect to manually constructed resources
    - Although potentially noisy
- Rely on freely available resource: Wiktionary
    - Large coverage and decent quality across many languages

Work packages



WP1

Corpus preparation, lexicon design, corpus-lexicon interface

WP2

Weakly supervised induction of lexical senses and semantic frames

WP3

Generation of human-readable descriptions for induced frames

WP4

Semantic lexicon at the service of interpretability

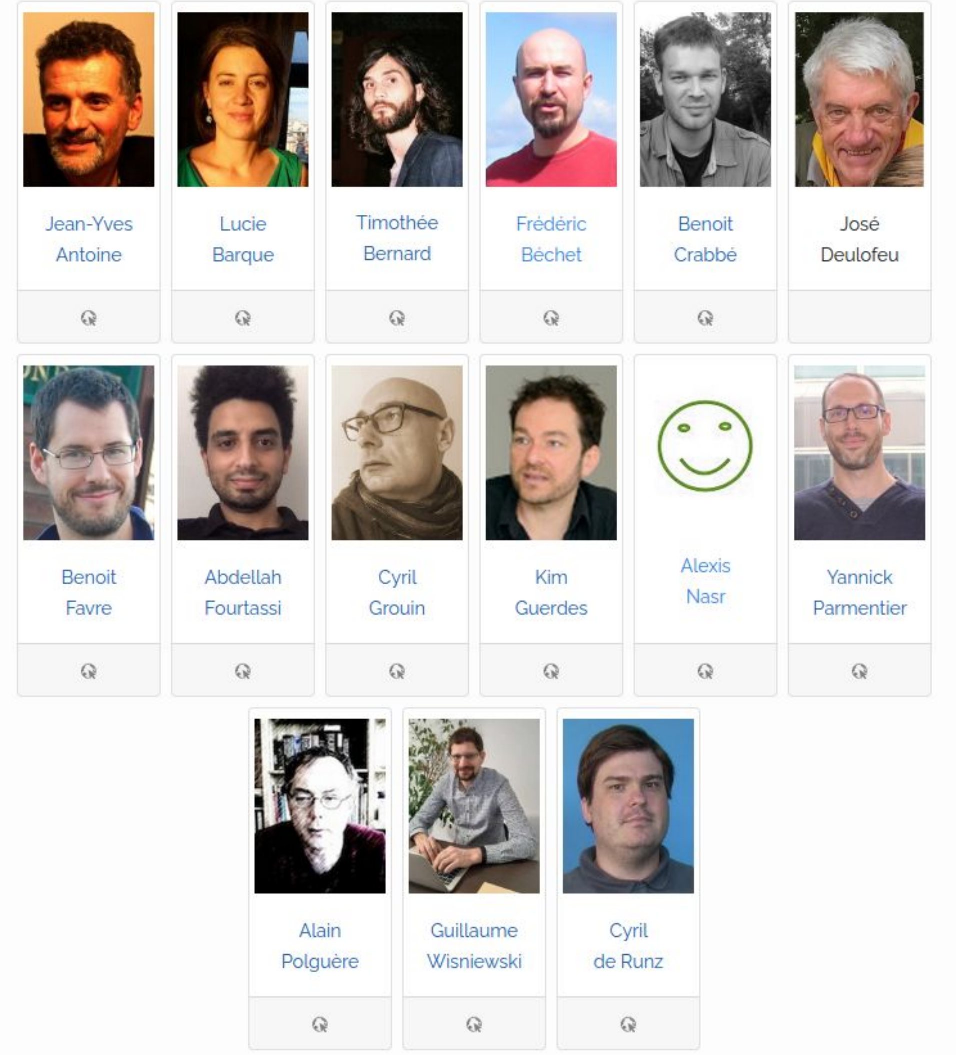
WP5

Semantic lexicon at the service of diversity

Consortium

- LIS - Aix Marseille Université (**C. Ramisch - PI**)
  LLF - Université de Paris (M. Candito)
- ATILF - CNRS Grand Est (M. Constant)
- LISN - Université de Paris-Saclay (A. Savary)
- LIFAT - Université de Tours (A. Soulet)

Participants



Jean-Yves Antoine

Lucie Barque

Timothée Bernard

Frédéric Béchet

Benoit Crabbé

José Deulofeu

Benoit Favre

Abdellah Fourtassi

Cyril Grouin

Kim Guerdes

Alexis Nasr

Yannick Parmentier

Alain Polguère

Guillaume Wisniewski

Cyril de Runz

Project infrastructure (WP0)

- Website
  - https://selexini.lis-lab.fr/
- Mailing lists
  - Selexini-all@lisn.upsaclay.fr
  - Selexini-core@lisn.upsaclay.fr
- Processing node
  - Part of LIS cluster, node "selexini-1"
  - 2 GPU Nvidia A100-80GB
  - Access upon request (create invited LIS account)
  - Priority for jobs of project members
- Comics version
  - Work in progress in collaboration with artist Marion Cluzel
- TODO
  - Logo suggestions
  - Project management (gitlab, agile tools…)
  - Social media presence

People

- Engineer WP1 - Marseille
    - Tithir Kumar Saha
- PhD thesis WP2 - Paris
    - Anna Mosolova
- Post-doc WP3 - Nancy
- Phd thesis WP4 - Marseille
- Post-doc WP5 - Saclay/Blois
- Internships
    - Wiktionary-based WSD for French: Ioana Ivan & Nathan Chometton - Marseille
    - …

Tour de table

A few words on your background?

What are your research interests?

Why are you here? What do you find interesting in SELEXINI?

How would you like to make a contribution (if any)?