



**chist-era**



# Year I of Interactive Grounded Language Understanding

Jean Rouat & Simon Brodeur, E&C Engineering  
Dept. (GEGI)

Neurocomputational and Intelligent Signal  
Processing Research Group (NECOTIS)

<https://www.gel.usherbrooke.ca/necotis/>  
Projects Seminar Meeting, 21 March 2017,  
Bruxelles



**NECOTIS**

neurosciences computationnelles et traitement  
intelligent des signaux



UNIVERSITÉ DE  
**SHERBROOKE**

# Who is contributing to the project?

- 8 research teams, across 6 different countries, it is a total effort of 325 person-months

- Experts:

Deep learning: A. Courville

Reinforcement learning: B. Piot, O. Pietquin (O. Colot)

Neurosciences & cognitive sciences: J. Rouat, R.K. Moore

Robotics: A.C. Murillo, J. Civera, M. Lopes (P.Y. Oudeyer)

Signal Processing: S. Dupont, G. Salvi, J. Rouat

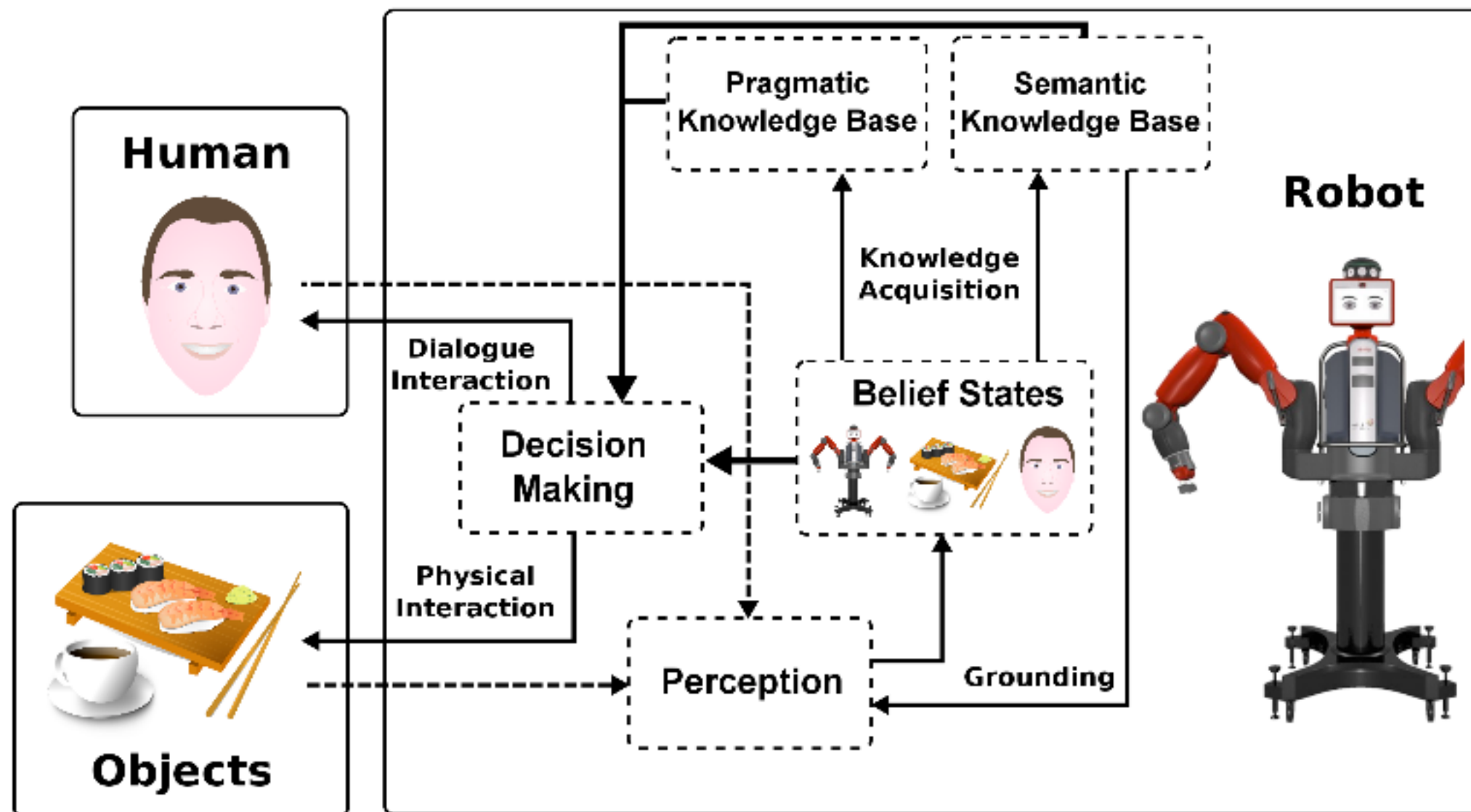
Human-machine Interaction: G. Salvi, S. Dupont

- 11 graduate students (5 in Québec)



*J. Rouat, Projects Seminar Meeting, 21 March 2017, Bruxelles*

# Unified Architecture



# Objectives

- Contribute to HLU with a unification of the respective contributions into a single architecture:
  - Semantic knowledge acquisition, modelling and grounding;
  - Pragmatic knowledge acquisition through interaction and modelling;
  - Decision-making based on reinforcement learning able to tackle semantic and pragmatic knowledge representations;
- Use the architecture in the context of human-agent interaction, aiming for human language acquisition and understanding to stimulate original research.
- Human Language is very broad (not only speech), in this context, ASR is another « textual labelling » modality for the IGLU architecture



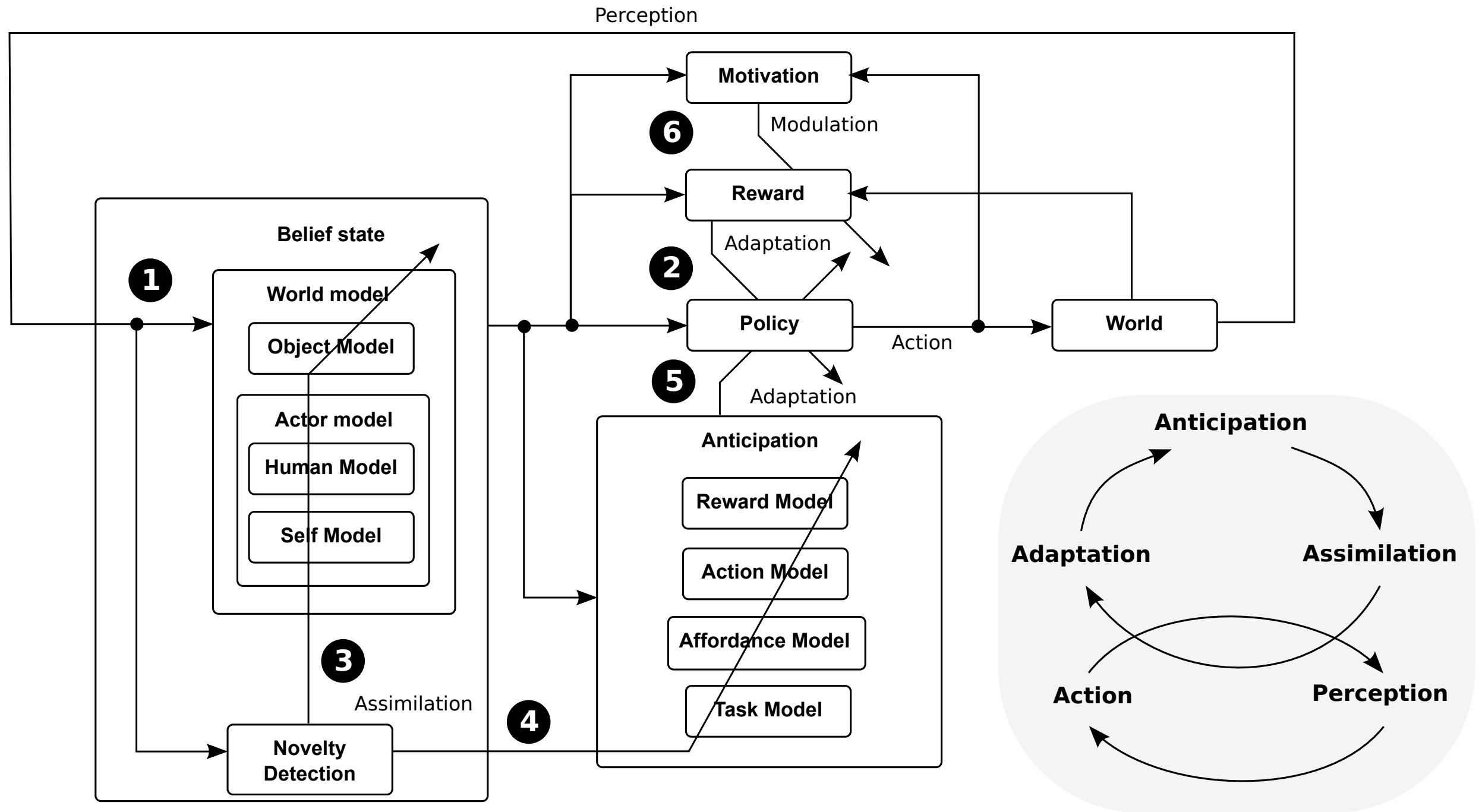
# Year I Summary

- A first cognitive architecture prototype, with a preliminary study of objective evaluation criteria and measures.
- Initial studies on novelty detection, a sound source separation system based on GCC-NMF, a neural network architecture that comprises an auto-encoder for speech recognition, a proposal for incremental learning of multimodal objects and an interactive dense segmentation of images to be used for databases labelling.
- 3 new databases that cover different levels of knowledge types and representations yielding a gradation in semantic representation and levels of interactions and grounding.
- A preliminary study on the use of game strategies (Nash equilibrium) for human-robot interaction, with study of potential links between language understanding.



*J. Rouat, Projects Seminar Meeting, 21 March 2017, Bruxelles*

# Cognitive Architecture Design and Evaluation





# Dataset GuessWhat?!

**Topic:** question answering with visual search

**Type:** supervised learning on a large dataset

**Evaluation:** perception (visual only), action (dialogue only)

**Scenario:** a cooperative two-player guessing game to locate an unknown object in a rich image scene



Is it on the front?	Yes
Is it a person?	No
Is it on the left?	No
Is it a snowboard?	Yes



Is it a cow?	Yes
Is it on the right?	No
Is it on the front?	No
Is it standing alone?	Yes



Is it one of the boys?	No
Is a boy carrying it?	Yes
Is it blue?	No
Is it a bag?	Yes

## Statistics:

- ▶ 155,280 played games, 821,889 question-answer pairs acquired with Amazon Mechanical Turk
- ▶ 66,537 images and 134,073 objects derived from the MS Coco dataset

**Link:** <https://guesswhat.ai/>



*J. Rouat, Projects Seminar Meeting, 21 March 2017, Bruxelles*



# Dataset Multimodal Human Robot Interaction

**Topic:** visual object and language learning in interaction

**Type:** unsupervised online learning on a small dataset

**Evaluation:** perception (visual only), adaptation, assimilation

**Scenarios:** the user interacts with the (passive) robot



Pointing to an object



Showing the object

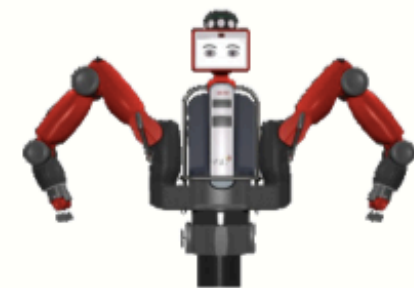


Talking about an object

## Statistics:

- ▶ 10 users with 30 interactions per user
- ▶ 22 common kitchen objects
- ▶ Multimodal data with RGB-D, RGB and audio

**Link:** <http://robots.unizar.es/IGLUdataset/>



*J. Rouat, Projects Seminar Meeting, 21 March 2017, Bruxelles*



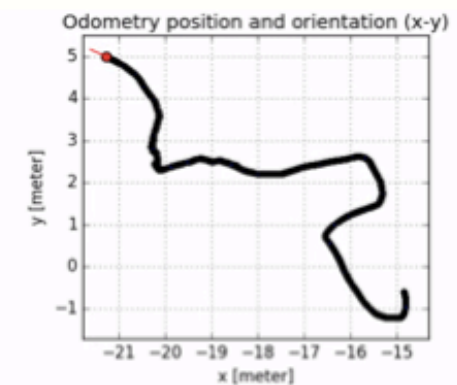
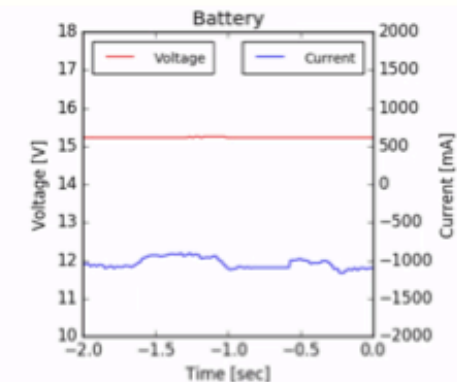
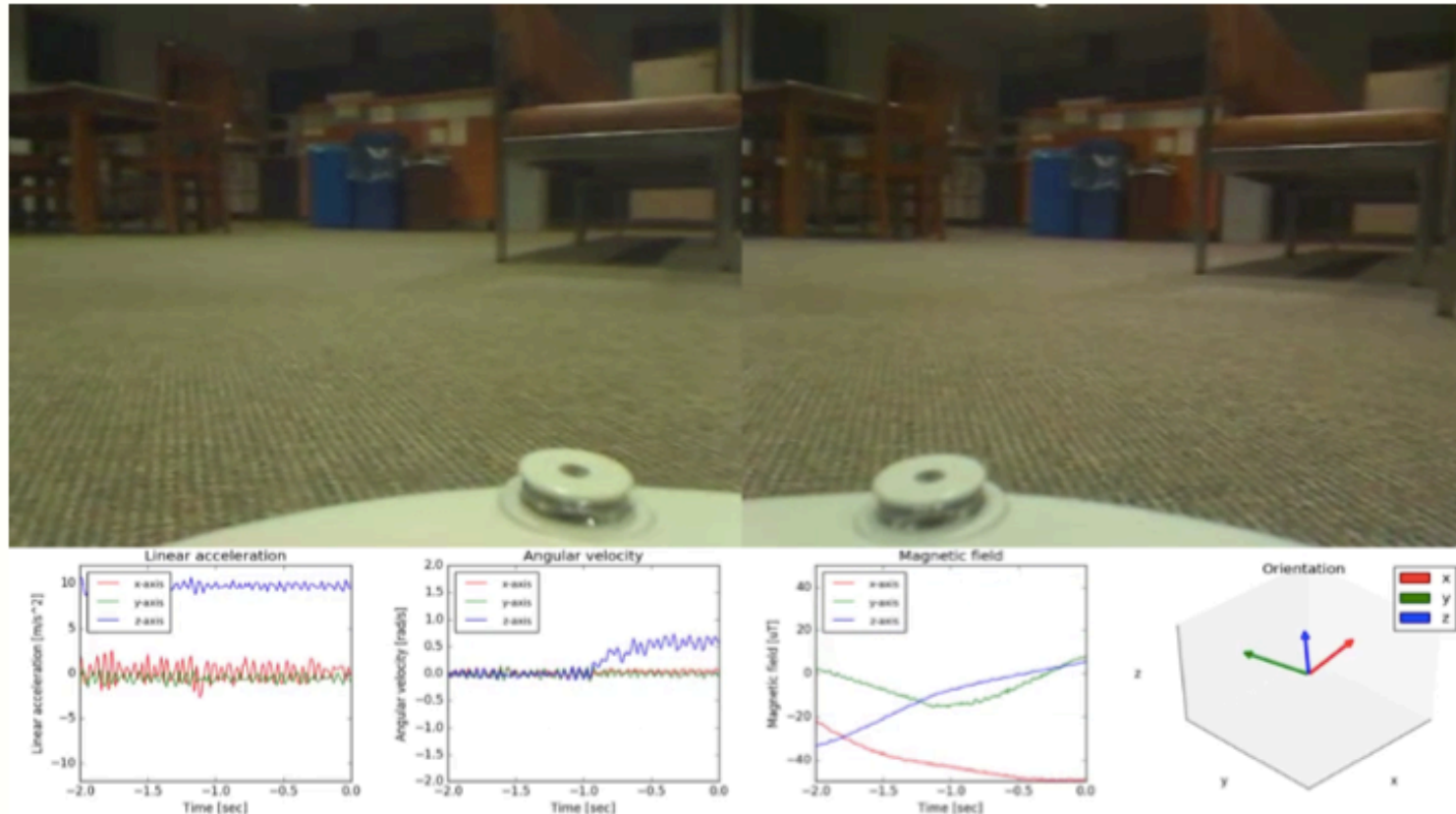
# Dataset Create

**Topic:** multimodal feature learning and novelty detection

**Type:** unsupervised online learning on a small dataset

**Evaluation:** perception (multimodal), adaptation, assimilation, anticipation

**Scenarios:** the robot explores and interacts with its environment



## Statistics:

- ▶ Wandering in environment in 8 rooms (4 trials each)
- ▶ Observing people in 4 rooms (4 trials each)
- ▶ Interacting with human in 1 room (8 trials each)
- ▶ 10 hours of total multimodal data

**Link:** <https://github.com/sbrodeur/ros-icreate-bbb>



*J. Rouat, Projects Seminar Meeting, 21 March 2017, Bruxelles*

# Publications & <https://iglu-chistera.github.io>

- A. Pablo, Y. Mollard, F. Golemo, A. C. Murillo, M. Lopes and J. Civera, "A Multimodal Human-Robot Interaction Dataset", Future of Interactive Learning Machines Workshop, NIPS 2016. Barcelona, Spain.
- A. B. Cambra, A. Muñoz, J. J. Guerrero and A. C. Murillo, "Dense Labeling with User Interaction: An Example for Depth-Of-Field Simulation", British Machine Vision Conference (BMVC), 2016.
- P. Azagra, F. Golemo, Y. Mollard, A. C. Murillo and J. Civera, "A Multimodal Dataset for Object Model Learning from Natural Human-Robot Interaction", (submitted to IROS 2017).
- J. Pérolat, F. Strub, B. Piot and O. Pietquin, "Learning Nash Equilibrium for General-Sum Markov Games from Batch Data", arXiv preprint arXiv:1606.08718, Accepted at the International Conference on Artificial Intelligence and Statistics 2017 (AISTAT 2017).
- H. de Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle and A. Courville, "GuessWhat?! Visual object discovery through multi-modal dialogue", arXiv preprint arXiv:1611.08481 (submitted to CVPR 2017).
- S. U. Wood, J. Rouat, S. Dupont and G. Pironkov, "Blind Speech Separation and Enhancement with GCC-NMF", IEEE Transactions on Audio, Speech and Language Processing, pp. 3329-3341, 2017.
- A. Dhaka and G. Salvi, "Semi-Supervised Learning with Sparse Autoencoders in Phone Classification", (submitted to INTERSPEECH



*J. Rouat, Projects Seminar Meeting, 21 March 2017, Bruxelles*



# Workshop IGLU



## International Workshop on Grounding Language Understanding, Satellite of Interspeech 2017

The purpose of this workshop is to bring together researchers working with modeling different aspects of language acquisition and understanding.

Although speech and language technology have reached a level of maturity, machines still fall short of human performance, especially when considering flexibility and robustness. It is therefore desirable to extend the machine learning approach which has been applied to speech technology, and emulate more closely the way humans acquire language. Desirable properties of the new learning approaches would include:

- being less dependent on linguistically annotated data,
- acquiring knowledge from multimodal inputs (acoustic, visual, tactile),
- learning from interaction,
- relating learning to the situational context,
- grounding language in the perceptual, emotional and sensorimotor experience of the system

This requires a highly multidisciplinary approach, combining the field of human language processing, speech technology, machine learning, developmental robotics and cognitive

### Important Dates

Submission deadline: Wednesday, 24 May 2017.

Workshop date: Friday, 25 August 2017.

### Endorsement

Official **INTERSPEECH 2017** Satellite Event

Supported by the **International Speech Communication Association**



*J. Rouat, Projects Seminar Meeting, 21 March 2017, Bruxelles*

# Year 2

- Continue our inter-team collaborations with student's exchange (Lille <-> MTL; UMONS <-> Sherbrooke; Bordeaux <-> Zaragoza, etc.).
- We are setting up a new team on Unsupervised Learning from Videos and Spoken Description (KTH, UMONS, UdeM & UdeS).
- Also a new team on affordances of the objects given their own actions and the other agent's actions (KTH, INRIA, UMONS).
- Words and affordances to learn associations between words, object properties, actions and effects in robot manipulation scenario (KTH, UNIZAR, INRIA, UMONS?, UdeS?).



# Expected Impacts

- Scientific impacts on machine learning and knowledge representation
  - Move toward interaction and cooperation with situated agent, where temporal aspect is important.
  - Learning on spatio-temporal multimodal data in relation with lifelong learning could be the next leap leading to even more success in complex problems.
  - Data and semantic knowledge collected from the project can contribute to the scientific effort.
  - How to model high-level, semantic and pragmatic knowledge in a robust fashion.


# Expected Impacts

- Scientific impacts on neuroscience and cognitive science
  - Emerging trend in neurosciences to consider the brain as a hierarchical generative model of the world.
  - New highlights and contribution to the parallel scientific debate of connectionism versus symbolism for cognition.
- Long-term societal impacts
  - Bring robotic agents in closer cooperation on daily tasks by having good interpretation and communication skills, as well as extensible knowledge.
  - Intelligent robotic agents of the future would adapt naturally to the user, based on the (perceived and simulated) experience of the user.

# Conclusion

- Followed the initial planning except that we decided not to be limited to the robot cooking application.
- A more general framework to be more or less independent of the cooking application.
- We still keep in mind that one application will be in the field of cooking robots but not the only one.
- With this strategy in mind, we expect a broader impact of IGLU in terms of research while not being exclusive to cooking robots.
- Because of delays in KTH, UMONS & UNIZAR, human model interaction has been a little bit delayed without impacts on the overall project.

# For more details: poster, tomorrow




## Year 1 of Interactive Grounded Language Understanding

Simon Brodeur<sup>1</sup>, Mathilde Brousmiche<sup>1,3</sup>, Huseyin Cakmak<sup>3</sup>, Luca Celotti<sup>1</sup>, Javier Civera<sup>6</sup>, Aaron Courville<sup>5</sup>, Harm de Vries<sup>5</sup>, Stéphane Dupont<sup>3</sup>, Florian Golemo<sup>7</sup>, Manuel Lopes<sup>7</sup>, Olivier Mastropietro<sup>5</sup>, Pablo A. Millan<sup>6</sup>, Yoan Mollard<sup>7</sup>, Roger K. Moore<sup>8</sup>, Ana C. Murillo<sup>6</sup>, Pierre-Yves Oudeyer<sup>7</sup>, Olivier Pietquin<sup>2</sup>, Bilal Piot<sup>2</sup>, Gueorgui Pironkov<sup>3</sup>, Jean Rouat<sup>1</sup>, Giampiero Salvi<sup>4</sup>, Florian Strub<sup>2</sup>, Sean Wood<sup>1</sup>

<sup>1</sup> Université de Sherbrooke, Canada  
<sup>4</sup> Royal Institute of Technology, Sweden

<sup>2</sup> Université de Lille 1, France  
<sup>5</sup> Université de Montréal, Canada  
<sup>7</sup> Inria Bordeaux Sud-Ouest, France

<sup>3</sup> Université de Mons, Belgium  
<sup>6</sup> Universidad de Zaragoza, Spain  
<sup>8</sup> University of Sheffield, United-Kingdom



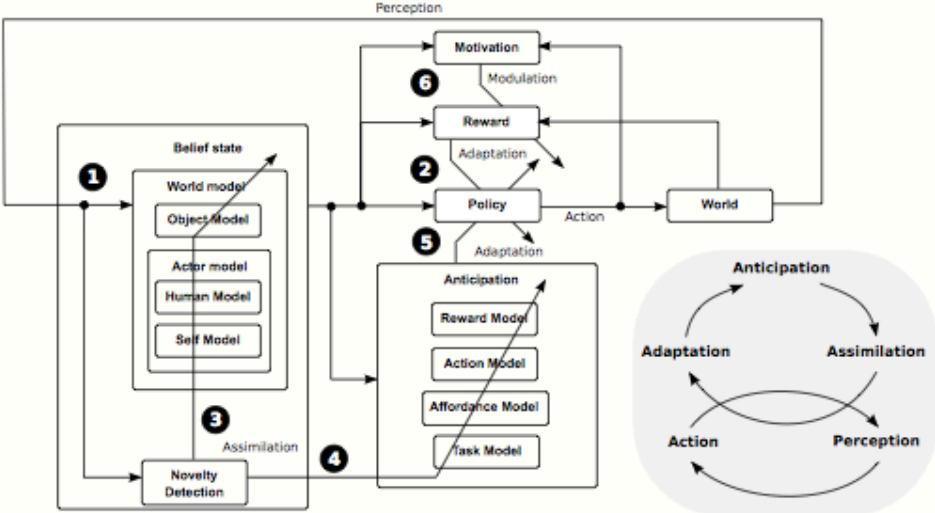
Authors by alphabetical order

### Summary

- First cognitive architecture, with study of objective evaluation criteria and measures.
- Initial study on novelty detection, a sound source separation system based on GCC-NMF, a neural network architecture that comprises an auto-encoder for speech recognition, a proposal for incremental learning of multimodal objects and an interactive dense segmentation of images to be used for databases labelling.
- 3 databases recorded that cover different levels of knowledge types and representations yielding a gradation in semantic representation and levels of interactions and grounding.
- Preliminary study on the use of game strategies (Nash equilibrium) for human-robot interaction, with study of potential links between language understanding.

### Cognitive Architecture Design and Evaluation

**Goal architecture:**

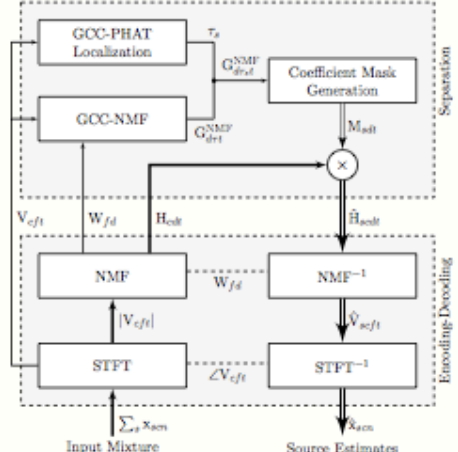


**Perform several tasks:**

- Extract a representation of the state of the environment
- Learn the optimal policy based on the reward function
- Detect novelty and adapt models for better perception
- Detect novelty and adapt models for better anticipation
- Adapt the policy to account for predictions
- Adapt the reward function to account for motivations

### Algorithm-related work

#### Blind Speech Separation and Enhancement with GCC-NMF



Input Mixture  $\sum_{s=1}^S x_{s,m}$  is processed by STFT to produce  $V_{cft}$ . This is then processed by NMF to produce  $W_{fd}$  and  $H_{cft}$ .  $H_{cft}$  is processed by NMF-1 to produce  $\hat{V}_{cft}$ , which is then processed by STFT-1 to produce Source Estimates  $\hat{x}_{s,m}$ . The coefficient mask generation block takes  $H_{cft}$  and  $\hat{V}_{cft}$  to produce  $M_{cft}$ , which is then used to process  $V_{cft}$  to produce  $\hat{V}_{cft}$ .

#### Semi-Supervised Learning with Sparse Autoencoders

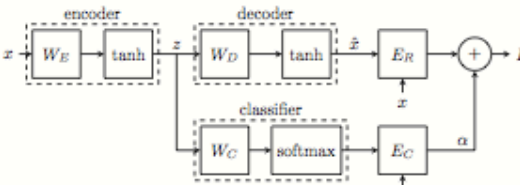
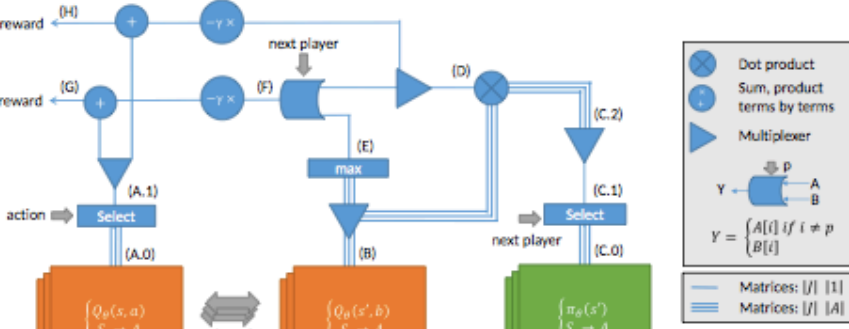


Figure: Architecture of the semi-supervised method for phoneme classification. The weights of the encoder, decoder and classifier are learned with a combination of labelled and unlabelled data. For unlabelled data, the reconstruction loss  $E_R$  is optimized, whereas the cross-entropy  $E_C$  loss is optimized over labelled data.

#### Learning Nash Equilibrium for General-Sum Markov Games from Batch Data



Legend:
 

- Dot product:  $\odot$
- Sum, product terms by terms:  $\oplus$
- Multiplexer:  $\triangleright$
- Selection:  $\text{Select}$
- Matrices:  $[I] \quad [1]$
- Matrices:  $[I] \quad [A]$



J. Rouat, Projects Seminar Meeting, 21 March 2017, Bruxelles



# Web site



*J. Rouat, Projects Seminar Meeting, 21 March 2017, Bruxelles*