Activity Report 2019

Team COML

Cognitive Machine Learning

Inria teams are typically groups of researchers working on the definition of a common project, and objectives, with the goal to arrive at the creation of a project-team. Such project-teams may include other partners (universities or research institutions).

RESEARCH CENTER
Paris

THEME
Language, Speech and Audio

Table of contents

Team COML

Creation of the Team: 2017 May 04

Keywords:

Computer Science and Digital Science:

A2.5.1. - Software Architecture & Design
A2.5.4. - Software Maintenance & Evolution
A2.5.5. - Software testing
A3.4.2. - Unsupervised learning
A3.4.5. - Bayesian methods
A3.4.6. - Neural networks
A3.4.8. - Deep learning
A5.7. - Audio modeling and processing
A5.7.1. - Sound
A5.7.3. - Speech
A5.7.4. - Analysis
A5.8. - Natural language processing
A6.3.3. - Data processing
A9.2. - Machine learning
A9.3. - Signal analysis
A9.4. - Natural language processing
A9.7. - AI algorithmics

Other Research Topics and Application Domains:

B1.2. - Neuroscience and cognitive science
B1.2.2. - Cognitive science
B2.2.6. - Neurodegenerative diseases
B2.5.2. - Cognitive disabilities
B9.6.1. - Psychology
B9.6.8. - Linguistics
B9.8. - Reproducibility
B9.10. - Privacy

1. Team, Visitors, External Collaborators

Research Scientist

Justine Cassell [Inria, Advanced Research Position, from Oct 2019]

Faculty Member

Emmanuel Dupoux [Team leader, École des hautes études en sciences sociales, Professor, HDR]

PhD Students

Robin Algayres [École Normale Supérieure de Paris, PhD Student, from Oct 2019]
Rahma Chaabouni [École Normale Supérieure de Paris, PhD Student]
Juliette Millet [École Normale Supérieure de Paris, PhD Student]
Rachid Riad [École Normale Supérieure de Paris, PhD Student]

Neil Zeghidour [Facebook, PhD Student, until Feb 2019]
Ronan Riochet [PhD Student, co-advised E. Dupoux, I. Laptev, J. Sivic]

Technical staff

Robin Algayres [École Normale Supérieure de Paris, Engineer, until Sep 2019]
Mathieu Bernard [Inria, Engineer]
Xuan-Nga Cao [École des hautes études en sciences sociales, Engineer]
Nicolas Hamilakis [École Normale Supérieure de Paris, Engineer]
Julien Karadayi [Inria, Engineer, from Apr 2019]
Manel Khentout [École Normale Supérieure de Paris, Engineer]
Marvin Lavechin [École Normale Supérieure de Paris, Engineer]
Malik Ould Arbi [CNRS, Engineer, from Feb 2019]
Hadrien Titeux [École Normale Supérieure de Paris, Engineer]
Kristina Madden [École Normale Supérieure de Paris, Designer]
Mohamed Zaiem [Univ Denis Diderot, Engineer, from Sep 2019]

Administrative Assistants

Chantal Chazelas [Inria, Administrative Assistant, until Jun 2019]
Meriem Guemair [Inria, Administrative Assistant, until Nov 2019]
Catherine Urban [École des hautes études en sciences sociales, Administrative Assistant]

External Collaborators

Nawal Abboub [École Normale Supérieure de Paris, from Apr 2019 until Jul 2019]
Ewan Dunbar [Univ Denis Diderot, from Apr 2019]

2. Overall Objectives

2.1. Overall Objectives

Brain-inspired machine learning algorithms combined with big data have recently reached spectacular results, equalling or beating humans on specific high level tasks (e.g. the game of go). However, there are still a lot of domains in which even humans infants outperform machines: unsupervised learning of rules and language, common sense reasoning, and more generally, cognitive flexibility (the ability to quickly transfer competence from one domain to another one).

The aim of the Cognitive Computing team is to *reverse engineer* such human abilities, i.e., to construct effective and scalable algorithms which perform as well (or better) than humans, when provided with similar data, study their mathematical and algorithmic properties and test their empirical validity as models of humans by comparing their output with behavioral and neuroscientific data. The expected results are more adaptable and autonomous machine learning algorithm for complex tasks, and quantitative models of cognitive processes which can used to predict human developmental and processing data. Most of the work is focused on speech and language and common sense reasoning.

3. Research Program

3.1. Background

In recent years, Artificial Intelligence (AI) has achieved important landmarks in matching or surpassing human level performance on a number of high level tasks (playing chess and go, driving cars, categorizing picture, etc., [31], [34], [39], [30], [36]). These strong advances were obtained by deploying on large amounts of data, massively parallel learning architectures with simple brain-inspired 'neuronal' elements. However, humans brains still outperform machines in several key areas (language, social interactions, common sense reasoning, motor skills), and are more flexible : Whereas machines require extensive expert knowledge and

massive training for each particular application, humans learn autonomously over several time scales: over the developmental scale (months), humans infants acquire cognitive skills with noisy data and little or no expert feedback (weakly/unsupervised learning)[1]; over the short time scale (minutes, seconds), humans combine previously acquired skills to solve new tasks and apply rules systematically to draw inferences on the basis of extremely scarce data (learning to learn, domain adaptation, one- or zero-shot learning) [33].

The general aim of CoML, following the roadmap described in [1], is to bridge the gap in cognitive flexibility between humans and machines learning in language processing and common sense reasoning by reverse engineering how young children between 1 and 4 years of age learn from their environment. We conduct work along two axes: the first one, which we called *Developmental AI* is focused on building infant inspired machine learning algorithms. The second axis is devoted to using the developed algorithms to conduct *quantitative studies* of how infant learn across diverse environments.

3.2. Weakly/Unsupervised Learning

Much of standard machine learning is construed as regression or classification problems (mapping input data to expert-provided labels). Human infants rarely learn in this fashion, at least before going to school: they learn language, social cognition, and common sense autonomously (without expert labels) and when adults provide feedback, it is ambiguous and noisy and cannot be taken as a gold standard. Modeling or mimicking such achievement requires deploying unsupervised or weakly supervised algorithms which are less well known than their supervised counterparts.

We take inspiration from infant's landmarks during their first years of life: they are able to learn acoustic models, a lexicon, and susbtantive elements of language models and world models from raw sensory inputs. Building on previous work [3], [7], [11], we use DNN and Bayesian architectures to model the emergence of linguistic representations without supervision. Our focus is to establish how the labels in supervised settings can be replaced by weaker signals coming either from multi-modal input or from hierarchically organised linguistic levels.

At the level of phonetic representations, we study how cross-modal information (lips and self feedback from articulation) can supplement top-down lexical information in a weakly supervised setting. We use Siamese architectures or Deep CCA algorithms to combine the different views. We study how an attentional framework and uncertainty estimation can flexibly combine these informations in order to adapt to situations where one view is selectively degraded.

At the level of lexical representations, we study how audio/visual parallel information (ie. descriptions of images or activities) can help in segmenting and clustering word forms, and vice versa, help in deriving useful visual features. To achieve this, we will use architectures deployed in image captioning or sequence to sequence translation [37].

At the level of semantic and conceptual representations, we study how it is possible to learn elements of the laws of physics through the observation of videos (object permanence, solidity, spatio-temporal continuity, inertia, etc.), and how objects and relations between objects are mapped onto language.

3.3. Evaluating Machine Intelligence

Increasingly, complicated machine learning systems are being incorporated into real-life applications (e.g. self-driving cars, personal assistants), even though they cannot be formally verified, guaranteed statistically, nor even explained. In these cases, a well defined *empirical approach* to evaluation can offer interesting insights into the functioning and offer some control over these algorithms.

Several approaches exist to evaluate the 'cognitive' abilities of machines, from the subjective comparison of human and machine performance [38] to application-specific metrics (e.g., in speech, word error rate). A recent idea consist in evaluating an AI system in terms of it's *abilities* [32] , i.e., functional components within a more global cognitive architecture [35]. Psychophysical testing can offer batteries of tests using simple tasks that are easy to understand by humans or animals (e.g, judging whether two stimuli are same or different, or

judging whether one stimulus is 'typical') which can be made selective to a specific component and to rare but difficult or adversarial cases. Evaluations of learning rate, domain adaptation and transfer learning are simple applications of these measures. Psychophysically inspired tests have been proposed for unsupervised speech and language learning [10], [6].

3.4. Documenting human learning

Infants learn their first language in a spontaneous fashion, across a lot of variation in amount of speech and the nature of the infant/adult interaction. In some linguistic communities, adults barely address infants until they can themselves speak. Despite these large variations in quantity and content, language learning proceeds at similar paces. Documenting such resilience is an essential step in understanding the nature of the learning algorithms used by human infants. Hence, we propose to collect and/or analyse large datasets of inputs to infants and correlate this with outcome measure (phonetic learning, vocabulary growth, syntactic learning, etc.).

4. Application Domains

4.1. Speech processing for underresourced languages

We plan to apply our algorithms for the unsupervised discovery of speech units to problems relevant to language documentation and the construction of speech processing pipelines for underresourced languages.

4.2. Tools for the analysis of naturalistic speech corpora

Daylong recordings of speech in the wild gives rise a to number of specific analysis difficulties. We plan to use our expertise in speech processing to develop tools for performing signal processing and helping annotation of such resources for the purpose of phonetic or linguistic analysis.

5. New Software and Platforms

5.1. intphys

IntPhys: A Benchmark for Visual Intuitive Physics Reasoning

KEYWORDS: Competition - Physical simulation - Artificial intelligence - Video Game

FUNCTIONAL DESCRIPTION: The intphys benchmark can be applied to any vision system, engineered, or trained, provided it can output a scalar when presented with a video clip, which should correspond to how physically plausible the video clip is. Our test set contains well matched videos of possible versus impossible events, and the metric consists in measuring how well the vision system can tell apart the possible from the impossible events..

- Contact: Mathieu Bernard
- URL: http://www.intphys.com

5.2. shennong

KEYWORDS: Speech processing - Python - Information extraction - Audio signal processing

FUNCTIONAL DESCRIPTION: Shennong is a Python library which implement the most used methods for speech features extraction. Features extraction is the first step of every speech processing pipeline.

Shennong provides the following functionalities: - implementation of the main methods from state of the art (including pre and post processing) - exhaustive documentation and tests - usage from a Python API or a command line tool - simple and coherent interface

- Contact: Mathieu Bernard
- URL: https://coml.lscp.ens.fr/docs/shennong

5.3. Seshat

Seshat Audio Annotation Platform

KEYWORDS: Audio - Speech - Web Application - Speech-text alignment

FUNCTIONAL DESCRIPTION: A web application to ease audio annotation campaigns, while also enabling the campaign manager to ensure that all annotations stick to a predefined format.

- Partner: ENS Paris
- Contact: Hadrien Titeux
- URL: https://github.com/bootphon/seshat

5.4. pyGammaAgreement

KEYWORDS: Reliability - Measures

FUNCTIONAL DESCRIPTION: Python library for measuring inter and intra annotator reliability for annotation sequences

- Contact: Emmanuel Dupoux

5.5. phonemizer

KEYWORD: Text

FUNCTIONAL DESCRIPTION:

- Conversion of a text into its phonemic representation
- Wrapper on speech synthesis programs espeak and festival
- Contact: Mathieu Bernard
- URL: https://github.com/bootphon/phonemizer

6. New Results

6.1. Unsupervised learning

Huamns learn to speak and to perceive the world in a largely self-supervised fashion. Yet, most of machine learning is still devoted to supervised algorithms that rely on abundant quantities of human labelled data. We have used humans as sources of inspiration for developing 3 novel machine learning benchmarks in order to push the field towards self-supervised learning.

- In the Zero Resource Speech Challenge 2019 [19], presented as a special session at Interspeech 2019, we propose to build a speech synthesizer without any text or phonetic labels: hence, TTS without T (text-to-speech without text). We provide raw audio for a target voice in an unknown language (the Voice dataset), but no alignment, text or labels. Participants must discover subword units in an unsupervised way (using the Unit Discovery dataset) and align them to the voice recordings in a way that works best for the purpose of synthesizing novel utterances from novel speakers, similar to the target speaker's voice. We describe the metrics used for evaluation, a baseline system consisting of unsupervised subword unit discovery plus a standard TTS system, and a topline TTS using gold phoneme transcriptions. We present an overview of the 19 submitted systems from 11 teams and discuss the main results.

- In [27], we introduce a new collection of spoken English audio suitable for training speech recognition systems under limited or no supervision. It is derived from open-source audio books from the LibriVox project. It contains over 60K hours of audio, which is, to our knowledge, the largest freely-available corpus of speech. The audio has been segmented using voice activity detection and is tagged with SNR, speaker ID and genre descriptions. Additionally, we provide baseline systems and evaluation metrics working under three settings: (1) the zero resource/unsupervised setting (ABX), (2) the semi-supervised setting (PER, CER) and (3) the distant supervision setting (WER). Settings (2) and (3) use limited textual resources (10 minutes to 10 hours) aligned with the speech. Setting (3) uses large amounts of unaligned text. They are evaluated on the standard LibriSpeech dev and test sets for comparison with the supervised state-of-the-art.
- In order to reach human performance on complex visual tasks, artificial systems need to incorporate a significant amount of understanding of the world in terms of macroscopic objects, movements, forces, etc. Inspired by work on intuitive physics in infants, we propose in [28] an evaluation framework which diagnoses how much a given system understands about physics by testing whether it can tell apart well matched videos of possible versus impossible events. The test requires systems to compute a physical plausibility score over an entire video. It is free of bias and can test a range of specific physical reasoning skills. We then describe the first release of a benchmark dataset aimed at learning intuitive physics in an unsupervised way, using videos constructed with a game engine. We describe two Deep Neural Network baseline systems trained with a future frame prediction objective and tested on the possible versus impossible discrimination task. The analysis of their results compared to human data gives novel insights in the potentials and limitations of next frame prediction architectures. This benchmark is currently being used in the DARPA project Machine Common Sense.

6.2. Language emergence in communicative agents

In this relatively new research topic, which is currently the focus of Rahma Chaabouni's PhD thesis, we study the inductive biases of neural systems by presenting them with few or no data.

- In [18], we study LSTMs' biases with respect to "natural" word-order constraints. To this end, we train them to communicate about trajectories in a grid world, using an artificial language that reflect or violate various natural language trends, such as the tendency to avoid redundancy or to minimize long-distance dependencies. We measure the speed of individual learning and the generational stability of language patterns in an iterative learning setting. Our results show a mixed picture. If LSTMs are affected by some "natural" word-order constraints, such as a preference for iconic orders and short-distance constructions, they have a preference toward redundant languages.
- In [25], we ask whether LSTMs have least-effort constraints and how this can affect their language. We let the neural systems develop their own language, to study a fundamental characteristic of natural language; Zipf's Law of Abbreviation (ZLA). In other words, we inverstigate if, even with the lack of the least-effort, LSTMs would produce a ZLA-like distribution like what we observe in natural language. Surprisingly, we find that networks develop an anti-efficient encoding scheme, in which the most frequent inputs are associated to the longest messages, and messages in general are skewed towards the maximum length threshold. This anti-efficient code appears easier to discriminate for the listener, and, unlike in human communication, the speaker does not impose a contrasting least-effort pressure towards brevity, as observed in [18]. Indeed, when the cost function includes a penalty for longer messages, the resulting message distribution starts respecting (ZLA). Our analysis stresses the importance of studying the basic features of emergent communication in a highly controlled setup, to ensure the latter will not strand too far from human language. Moreover, we present a concrete illustration of how different functional pressures can lead to successful communication codes that lack basic properties of human language, thus highlighting the role such pressures play in the latter.

- There is renewed interest in simulating language emergence among deep neural agents that communicate to jointly solve a task, spurred by the practical aim to develop language-enabled interactive AIs, and by theoretical questions about the evolution of human language. However, optimizing deep architectures connected by a discrete communication channel (such as that in which language emerges) is technically challenging. In [21], we introduce EGG, a toolkit that greatly simplifies the implementation of emergent-language communication experiments. EGG's modular design provides a set of building blocks that the user can combine to create new communication games, easily navigating the optimization and architecture space. We hope that the tool will lower the technical barrier, and encourage researchers from various backgrounds to do original work in this exciting area/

6.3. Evaluation of AI algorithms

Machine learning algorithms are typically evaluated in terms of end-to-end tasks, but it is very often difficult to get a grasp of how they achieve these tasks, what could be their break point, and more generally, how they would compare to the algorithms used by humans to do the same tasks. This is especially true of Deep Learning systems which are particularly opaque. The team develops evaluation methods based on psycholinguistic/linguistic criteria, and deploy them for systematic comparison of systems.

- Recurrent neural networks (RNNs) can learn continuous vector representations of symbolic structures such as sequences and sentences; these representations often exhibit linear regularities (analogies). Such regularities motivate our hypothesis that RNNs that show such regularities implicitly compile symbolic structures into tensor product representations (TPRs; Smolensky, 1990), which additively combine tensor products of vectors representing roles (e.g., sequence positions) and vectors representing fillers (e.g., particular words). To test this hypothesis, we introduce Tensor Product Decomposition Networks (TPDNs), which use TPRs to approximate existing vector representations. We demonstrate using synthetic data that TPDNs can successfully approximate linear and tree-based RNN autoencoder representations, suggesting that these representations exhibit interpretable compositional structure; we explore the settings that lead RNNs to induce such structure-sensitive representations. By contrast, further TPDN experiments show that the representations of four models trained to encode naturally-occurring sentences can be largely approximated with a bag of words, with only marginal improvements from more sophisticated structures. We conclude that TPDNs provide a powerful method for interpreting vector representations, and that standard RNNs can induce compositional sequence representations that are remarkably well approximated by TPRs; at the same time, existing training tasks for sentence representation learning may not be sufficient for inducing robust structural representations.
- LSTMs have proven very successful at language modeling. However, it remains unclear to what extent they are able to capture complex morphosyntactic structures. In [29], we examine whether LSTMs are sensitive to verb argument structures. We introduce a German grammaticality dataset in which ungrammatical sentences are constructed by manipulating case assignments (eg substituting nominative by accusative or dative). We find that LSTMs are better than chance in detecting incorrect argument structures and slightly worse than humans tested on the same dataset. Surprisingly, LSTMs are contaminated by heuristics not found in humans like a preference toward nominative noun phrases. In other respects they show human-similar results like biases for particular orders of case assignments.
- Pater (2019) proposes to use neural networks to model learning within existing grammatical frameworks. In [16] we argue that there is a fundamental gap to be bridged that does not receive enough attention : how can we use neural networks to examine whether it is possible to learn some linguistic representation (a tree, for example) when, after learning is finished, we cannot even tell if this is the type of representation that has been learned (all we see is a sequence of numbers)? Drawing a correspondence between an abstract linguistic representational system and an opaque parameter vector that can (or perhaps cannot) be seen as an instance of such a representation is an implementational mapping problem. Rather than relying on existing frameworks that propose

partial solutions to this problem, such as harmonic grammar, we suggest that fusional research of the kind proposed needs to directly address how to 'find' linguistic representations in neural network representations.

6.4. Learnability relevant descriptions of linguistic corpora

Evidently, infants are acquiring their language based on whatever linguistic input is available around them. The extent of variation that can be found across languages, cultures and socio-economic background provides strong constraints (lower bounds on data, higher bounds on noise, and variation and ambiguity) for language learning algorithms.

- Previous computational modeling suggests it is much easier to segment words from child-directed (CDS) than adult-directed speech (ADS). However, this conclusion is based on data collected in the laboratory, with CDS from play sessions and ADS between a parent and an experimenter, which may not be representative of ecologically-collected CDS and ADS. In [15], fully naturalistic ADS and CDS collected with a non-intrusive recording device as the child went about her day were analyzed with a diverse set of algorithms. The difference between registers was small compared to differences between algorithms, it reduced when corpora were matched, and it even reversed under some conditions. These results highlight the interest of studying learnability using naturalistic corpora and diverse algorithmic definitions.
- A number of unsupervised learning algorithms have been proposed in the last 20 years for modeling early word learning, some of which have been implemented computationally, but whose results remain difficult to compare across papers. In [14], we created a tool that is open source, enables reproducible results, and encourages cumulative science in this domain. WordSeg has a modular architecture: It combines a set of corpora description routines, multiple algorithms varying in complexity and cognitive assumptions (including several that were not publicly available, or insufficiently documented), and a rich evaluation package. In the paper, we illustrate the use of this package by analyzing a corpus of child-directed speech in various ways, which further allows us to make recommendations for experimental design of follow-up work. Supplementary materials allow readers to reproduce every result in this paper, and detailed online instructions further enable them to go beyond what we have done. Moreover, the system can be installed within container software that ensures a stable and reliable environment. Finally, by virtue of its modular architecture and transparency, WordSeg can work as an open-source platform, to which other researchers can add their own segmentation algorithms.

6.5. Test of the psychological validity of AI algorithms.

In this section, we focus on the utilisation of machine learning algorithms of speech and language processing to derive testable quantitative predictions in humans (adults or infants).

- In [24], we compare the performance of humans (English and French listeners) versus an unsupervised speech model in a perception experiment (ABX discrimination task). Although the ABX task has been used for acoustic model evaluation in previous research, the results have not, until now, been compared directly with human behaviour in an experiment. We show that a standard, well-performing model (DPGMM) has better accuracy at predicting human responses than the acoustic baseline. The model also shows a native language effect, better resembling native listeners of the language on which it was trained. However, the native language effect shown by the models is different than the one shown by the human listeners, and, notably , the models do not show the same overall patterns of vowel confusions.
- Word learning relies on the ability to master the sound con-trasts that are phonemic (i.e., signal meaning difference) in a given language. Though the timeline of phoneme development has been studied extensively over the past few decades, the mechanism of this development is poorly understood. In [20], we take inspiration from computational modeling work in language grounding

where phonetic and visual information is learned jointly. In this study, we varied the taxonomic distance of pairs of objects and tested how adult learners judged the phonemic status of the sound contrast associated with each of these pairs. We found that judgments were sensitive to gradients in the taxonomic structure, suggesting that learners use probabilistic information at the semantic level to optimize the accuracy of their judgements at the phonological level. The findings provide evidence for an interaction between phonological learning and meaning generalization in human learning.

6.6. Applications and tools for researchers

Some of CoMLs' activity is to produce speech and language technology tools that facilitate research into language development or clinical applications.

- Speech classifiers of paralinguistic traits traditionally learn from diverse hand-crafted low-level features, by selecting the relevant information for the task at hand. We explore an alternative to this selection, by learning jointly the classifier, and the feature extraction. Recent work on speech recognition has shown improved performance over speech features by learning from the waveform. In [24], we extend this approach to paralinguistic classification and propose a neural network that can learn a filterbank, a normalization factor and a compression power from the raw speech, jointly with the rest of the architecture. We apply this model to dysarthria detection from sentence-level audio recordings. Starting from a strong attention-based baseline on which mel-filterbanks out-perform standard low-level descriptors, we show that learning the filters or the normalization and compression improves over fixed features by 10% absolute accuracy. We also observe a gain over OpenSmile features by learning jointly the feature extraction, the normalization, and the compression factor with the architecture. This constitutes a first attempt at learning jointly all these operations from raw audio for a speech classification task.
- This paper [23] presents the problems and solutions addressed at the JSALT workshop when using a single microphone for speaker detection in adverse scenarios. The main focus was to tackle a wide range of conditions that go from meetings to wild speech. We describe the research threads we explored and a set of modules that was successful for these scenarios. The ultimate goal was to explore speaker detection; but our first finding was that an effective diarization improves detection, and not having a diarization stage impoverishes the performance. All the different configurations of our research agree on this fact and follow a main backbone that includes diarization as a previous stage. With this backbone, we analyzed the following problems: voice activity detection, how to deal with noisy signals, domain mismatch, how to improve the clustering; and the overall impact of previous stages in the final speaker detection. In this paper, we show partial results for speaker diarizarion to have a better understanding of the problem and we present the final results for speaker detection.

7. Bilateral Contracts and Grants with Industry

7.1. Bilateral Grants with Industry

- **Facebook AI Research Grant** (2019, PI: E. Dupoux, 350K€) - Unrestricted Gift - The aim is to help the development of machine learning tools geared towards the psycholinguistic research community.
- **Google Research Award** (2019, PI E. Dunbar, 37K€) - Unrestricted Gift - Develop a first version of a universal synthesizer which can be tuned to specific dialects with sparse data.

8. Partnerships and Cooperations

8.1. Regional Initiatives

Collaboration with the Willow Team:

- co-advising with J. Sivic and I. Laptev of a PhD student: Ronan Riochet.
- construction of a naive physics benchmark (http://www.intphys.com)

Collaboration with the Almanach Team:

- co-advising with B. Sagot a PhD student: Robin Algayres.
- co-advising with B. Sagot a Master student: Charlotte Rochereau

8.2. National Initiatives

8.2.1. ANR

- **ANR-Transatlantic Platform Digging into Data - ACLEW** (2017–2020. 5 countries; Total budget: 1.4M€; coordinating PI : M. Soderstrom; Local PI: A. Cristia; Leader of tools development and co-PI : E. Dupoux) - Constructing tools for the Analysis of Children's Language Experiences Around the World.
- **CNRS Prematuration - BabyCloud**. (2018-2019; coordinating PIs : E. Dupoux and X.-N. Cao; 100€) - Enable the construction of a fully functionnal Baby Logger prototype; perform a market analysis and prepare the launch of a startup.
- **ANR GEOMPHON**. (2018-2021; coordinating PI : E. Dunbar; 299K€) - Study the effects of typologically common properties of linguistic sound systems on speech perception, human learning, and machine learning applied to speech.

8.3. International Initiatives

8.3.1. Inria International Partners

8.3.1.1. Informal International Partners

- Johns Hopkins University, Baltimore, USA: S. Kudanpur, H. Hermanksy
- RIKEN Institute, Tokyo, Japan: R. Mazuka

8.4. International Research Visitors

8.4.1. Visits of International Scientists

Justine Cassell (CMU, ARP, PRAIRIE Chair starting from Oct 2019)

8.4.2. Visits to International Teams

8.4.2.1. Research Stays Abroad

+ E. Dupoux, Research Scientist, JSALT Workshop, Montreal (July, 2019)

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific Events Organisation

9.1.1.1. General Chair, Scientific Chair

- E. Dupoux & E. Dunbar: Co-organizers of the INTERSPEECH 2019 Zero Ressource Challenge Session.

9.1.1.2. Member of the Organizing Committees

- Executive committee of SIGMORPHON (Association for Computational Linguistics Special Interest Group, http://www.sigmorphon.org/).
- Executive committee of DARCLE http://www.darcle.org.

9.1.2. Scientific Events Selection

9.1.2.1. Reviewer

Invited editor for international conferences: Interspeech, NIPS, ACL, etc. (around 5-10 papers per conferences, 2 conferences per year)

9.1.3. Journal

9.1.3.1. Member of the Editorial Boards

Member of the editorial board of: Mathématiques et Sciences Humaines, L'Année Psychologique, Frontiers in Psychology.

9.1.3.2. Reviewer - Reviewing Activities

Invited Reviewer for Frontiers in Psychology, Cognitive Science, Cognition, Transactions in Acoustics Signal Processing and Language, Speech Communication, etc. (around 4 papers per year)

9.1.4. Invited Talks

- Dec/4/2019, E. Dupoux, Invited Speaker, Symposium NeuroDevRob: Developmental AI, Cergy Pontoise
- Jul/15/2019, E. Dupoux, Invited Seminar, Microsoft Research: Inductive Biases and Language Emergence in Communicative Agents, Seattle
- Jun/26/2019, E. Dupoux, Invited Speaker, DARPA MCS Kickoff Meeting: Measuring Intuitive Physics Understanding in Artificial Systems, Washington DC
- May/26/2019, E. Dupoux, Invited Speaker, Facebook CogSci-AI Workshop: Learning speech like infants do, New York
- Mar/14/2019, E. Dupoux, GDR TAL, Artificial models of language acquisition, Paris
- Jan/13/2019, E. Dupoux, Invited Seminar, UTC Compiegne, Developmental AI

9.1.5. Scientific Expertise

E. Dupoux is invited expert for ERC, ANR, and other granting agencies, or tenure committees (around 2 per year).

9.1.6. Research Administration

E. Dupoux is on the Executive committee of the Foundation Cognition, the research programme IRIS-PSL "Sciences des Données et Données des Sciences", the industrial chair Almerys (2016-) and the collective organization DARCLE (http://www.darcle.org).

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

E. Dupoux is co-director of the Cognitive Engineering track in the Cognitive Science Master (ENS, EHESS, Paris V).

- Master : E. Dupoux (with B. Sagot, ALMANACH, N. Zeghidour & R. Riad, COML), "Algorithms for speech and language processing", 30h, M2, (MVA), ENS Cachan, France
- Master : E. Dupoux, "Cognitive Engineering", 80h, M2, ITI-PSL, Paris France
- Doctorat : E. Dupoux, "Computational models of cognitive development", 32 h, Séminaire EHESS, Paris France
- Master: E. Dunbar, "Phonology" , 36 h, Master Sciences du Langage, Paris Diderot
- Master: E. Dunbar, "Statistics", 28h, Master Sciences du Langage, Paris Diderot
- Licence 3: E. Dunbar, "Phonology", 36h, Licence Sciences du Langage, Paris Diderot
- Licence 3: E. Dunbar, "Experimental methods", 36h, Licence Sciences du Langage, Paris Diderot

9.2.2. Supervision

- defended PhD: Neil Zeghidour, Learning speech features from raw signals, Feb 2015-13 Mar 2019, co-advised E. Dupoux, N. Usunier (Facebook-CIFRE)
- PhD in progress : Rahma Chaabouni, Language learning in artificial agents, Sept 2017, co-advised E. Dupoux, M. Baroni (Facebook-CIFRE)
- PhD in progress : Ronan Riochet, Learning models of intuitive physics, Sept 2017, co-advised E. Dupoux, I. Laptev, J. Sivic
- PhD in progress : Rachid Riad, "Speech technology for biomarkers in neurodegenerative diseases" , Sept 2018, co-advised E. Dupoux, A.-C. Bachoud-Levi
- PhD in progress : Robin Algayres "Audio word embeddings and word segmentation" , from Oct 2019, co-advised E. Dupoux, B. Sagot
- PhD in progress: Juliette Millet, "Modeling L2 Speech perception", from Sept 2018, advised E. Dunbar Bachoud-Lévi

9.2.3. Juries

E. Dupoux participated in the HDR jury of Jacques Cartier Jul 3, 2019.

9.3. Popularization

E. Dupoux presented the startup project BabyCloud at the CNRS innovatives SHS Salon, May 14-16, Lille.

10. Bibliography

Major publications by the team in recent years

[1] E. DUPOUX. *Cognitive Science in the era of Artificial Intelligence: A roadmap for reverse-engineering the infant language-learner*, in "Cognition", 2018

[2] A. FOURTASSI, E. DUPOUX. *A Rudimentary Lexicon and Semantics Help Bootstrap Phoneme Acquisition*, in "Proceedings of the 18th Conference on Computational Natural Language Learning (CoNLL)", Baltimore, Maryland USA, Association for Computational Linguistics, June 2014, pp. 191-200 [*DOI :* 10.3115/v1/W14-1620]

[3] A. FOURTASSI, T. SCHATZ, B. VARADARAJAN, E. DUPOUX. *Exploring the Relative Role of Bottom-up and Top-down Information in Phoneme Learning*, in "Proceedings of the 52nd Annual meeting of the ACL", Baltimore, Maryland, Association for Computational Linguistics, 2014, vol. 2, pp. 1-6 [*DOI :* 10.3115/v1/P14-2001]

[4] Y. HOSHEN, R. J. WEISS, K. W. WILSON. *Speech acoustic modeling from raw multichannel waveforms*, in "Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on", IEEE, 2015, pp. 4624–4628

[5] T. LINZEN, E. DUPOUX, Y. GOLDBERG. *Assessing the ability of LSTMs to learn syntax-sensitive dependencies*, in "Transactions of the Association for Computational Linguistics", 2016, vol. 4, pp. 521-535

[6] T. LINZEN, E. DUPOUX, B. SPECTOR. *Quantificational features in distributional word representations*, in "Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics", 2016, pp. 1-11 [*DOI :* 10.18653/v1/S16-2001]

[7] A. MARTIN, S. PEPERKAMP, E. DUPOUX. *Learning Phonemes with a Proto-lexicon*, in "Cognitive Science", 2013, vol. 37, pp. 103-124 [*DOI :* 10.1111/J.1551-6709.2012.01267.X]

[8] S. MEHRI, K. KUMAR, I. GULRAJANI, R. KUMAR, S. JAIN, J. SOTELO, A. COURVILLE, Y. BENGIO. *SampleRNN: An unconditional end-to-end neural audio generation model*, in "arXiv preprint arXiv:1612.07837", 2016

[9] T. N. SAINATH, R. J. WEISS, A. SENIOR, K. W. WILSON, O. VINYALS. *Learning the speech front-end with raw waveform CLDNNs*, in "Sixteenth Annual Conference of the International Speech Communication Association", 2015

[10] T. SCHATZ, V. PEDDINTI, F. BACH, A. JANSEN, H. HYNEK, E. DUPOUX. *Evaluating speech features with the Minimal-Pair ABX task: Analysis of the classical MFC/PLP pipeline*, in "INTERSPEECH-2013", Lyon, France, International Speech Communication Association, 2013, pp. 1781-1785

[11] R. THIOLLIÈRE, E. DUNBAR, G. SYNNAEVE, M. VERSTEEGH, E. DUPOUX. *A Hybrid Dynamic Time Warping-Deep Neural Network Architecture for Unsupervised Acoustic Modeling*, in "INTERSPEECH-2015", 2015, pp. 3179-3183

[12] A. VAN DEN OORD, S. DIELEMAN, H. ZEN, K. SIMONYAN, O. VINYALS, A. GRAVES, N. KALCHBRENNER, A. SENIOR, K. KAVUKCUOGLU. *Wavenet: A generative model for raw audio*, in "CoRR abs/1609.03499", 2016

Publications of the year

Doctoral Dissertations and Habilitation Theses

[13] N. ZEGHIDOUR. *Learning representations of speech from the raw waveform*, PSL Research University, March 2019, https://tel.archives-ouvertes.fr/tel-02278616

Articles in International Peer-Reviewed Journals

[14] M. BERNARD, R. THIOLLIÈRE, A. SAKSIDA, G. LOUKATOU, E. LARSEN, M. C. JOHNSON, L. FIBLA, E. DUPOUX, R. DALAND, X.-N. CAO, A. CRISTIA. *WordSeg: Standardizing unsupervised word form segmentation from text*, in "Behavior Research Methods", April 2019 [*DOI :* 10.3758/S13428-019-01223-3], https://hal.archives-ouvertes.fr/hal-02274072

[15] A. CRISTIA, E. DUPOUX, N. BERNSTEIN RATNER, M. SODERSTROM. *Segmentability Differences Between Child-Directed and Adult-Directed Speech: A Systematic Test With an Ecologically Valid Corpus*, in "Open Mind", 2019, vol. 3, pp. 13-22 [*DOI :* 10.1162/OPMI_A_00022], https://hal.archives-ouvertes.fr/hal-02274050

[16] E. DUNBAR. *Generative grammar, neural networks, and the implementational mapping problem: Response to Pater*, in "Language", 2019, vol. 95, n° 1, pp. e87-e98 [*DOI :* 10.1353/LAN.2019.0013], https://hal.archives-ouvertes.fr/hal-02274522

[17] M. MALDONADO, E. DUNBAR, E. CHEMLA. *Mouse tracking as a window into decision making*, in "Behavior Research Methods", June 2019, vol. 51, n° 3, pp. 1085-1101 [*DOI :* 10.3758/S13428-018-01194-X], https://hal.archives-ouvertes.fr/hal-02274523

International Conferences with Proceedings

[18] R. CHAABOUNI, E. KHARITONOV, A. LAZARIC, E. DUPOUX, M. BARONI. *Word-order biases in deep-agent emergent communication*, in "ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics", Florence, Italy, July 2019, https://arxiv.org/abs/1905.12330 , https://hal.archives-ouvertes.fr/hal-02274157

[19] E. DUNBAR, R. ALGAYRES, J. KARADAYI, M. BERNARD, J. BENJUMEA, X.-N. CAO, L. MISKIC, C. DUGRAIN, L. ONDEL, A. W. BLACK, L. BESACIER, S. SAKTI, E. DUPOUX. *The Zero Resource Speech Challenge 2019: TTS without T*, in "Interspeech 2019 - 20th Annual Conference of the International Speech Communication Association", Graz, Austria, September 2019, https://hal.archives-ouvertes.fr/hal-02274112

[20] A. FOURTASSI, E. DUPOUX. *Phoneme learning is influenced by the taxonomic organization of the semantic referents*, in "Cognitive Science Society", Montreal, Canada, July 2019, https://hal.archives-ouvertes.fr/hal-02274093

[21] E. KHARITONOV, R. CHAABOUNI, D. BOUCHACOURT, M. BARONI. *EGG: a toolkit for research on Emergence of lanGuage in Games*, in "Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations", Hong Kong, China, November 2019, https://arxiv.org/abs/1907.00852 [*DOI :* 10.18653/v1/D19-3010], https://hal.archives-ouvertes.fr/hal-02274229

[22] R. T. MCCOY, T. LINZEN, E. DUNBAR, P. SMOLENSKY. *RNNs Implicitly Implement Tensor Product Representations*, in "ICLR 2019 - International Conference on Learning Representations", New Orleans, United States, May 2019, https://arxiv.org/abs/1812.08718 - Accepted to ICLR 2019, https://hal.archives-ouvertes.fr/hal-02274498

[23] J. MILLET, N. JUROV, E. DUNBAR. *Comparing unsupervised speech learning directly to human performance in speech perception*, in "CogSci 2019 - 41st Annual Meeting of Cognitive Science Society", Montréal, Canada, July 2019, https://hal.archives-ouvertes.fr/hal-02274499

[24] J. MILLET, N. ZEGHIDOUR. *Learning to detect dysarthria from raw speech*, in "ICASSP-2019 - IEEE International Conference on Acoustics, Speech and Signal Processing", Brighton, United Kingdom, May 2019, https://arxiv.org/abs/1811.11101 , https://hal.archives-ouvertes.fr/hal-02274504

Other Publications

[25] R. CHAABOUNI, E. KHARITONOV, E. DUPOUX, M. BARONI. *Anti-efficient encoding in emergent communication*, August 2019, https://arxiv.org/abs/1905.12561 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-02274205

[26] P. GARCÍA, J. VILLALBA, H. BREDIN, J. DU, D. CASTAN, A. CRISTIA, L. BULLOCK, L. GUO, K. OKABE, P. S. NIDADAVOLU, S. KATARIA, S. CHEN, L. GALMANT, M. LAVECHIN, L. SUN, M.-P. GILL, B. BEN-YAIR, S. ABDOLI, X. WANG, W. BOUAZIZ, H. TITEUX, E. DUPOUX, K. A. LEE, N. DEHAK. *Speaker detection in the wild: Lessons learned from JSALT 2019*, December 2019, https://arxiv.org/abs/1912.00938 - Submitted to ICASSP 2020, https://hal.archives-ouvertes.fr/hal-02417632

[27] J. KAHN, M. RIVIÈRE, W. ZHENG, E. KHARITONOV, Q. XU, P.-E. MAZARÉ, J. KARADAYI, V. LIPTCHINSKY, R. COLLOBERT, C. FUEGEN, T. LIKHOMANENKO, G. SYNNAEVE, A. JOULIN, A. MOHAMED, E.

DUPOUX. *Libri-Light: A Benchmark for ASR with Limited or No Supervision*, December 2019, https://arxiv.org/abs/1912.07875 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-02417621

[28] R. RIOCHET, M. Y. CASTRO, M. BERNARD, A. LERER, R. FERGUS, V. IZARD, E. DUPOUX. *IntPhys: A Benchmark for Visual Intuitive Physics Reasoning*, August 2019, https://arxiv.org/abs/1803.07616 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-02274273

[29] C. ROCHEREAU, B. SAGOT, E. DUPOUX. *Modeling German Verb Argument Structures: LSTMs vs. Humans*, December 2019, https://arxiv.org/abs/1912.00239 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-02417640

References in notes

[30] D. A. FERRUCCI. *Introduction to "this is watson"*, in "IBM Journal of Research and Development", 2012, vol. 56, n^{o} 3.4, pp. 1–1

[31] K. HE, X. ZHANG, S. REN, J. SUN. *Delving deep into rectifiers: Surpassing human-level performance on imagenet classification*, in "Proceedings of the IEEE International Conference on Computer Vision", 2015, pp. 1026–1034

[32] J. HERNÁNDEZ-ORALLO, F. MARTÍNEZ-PLUMED, U. SCHMID, M. SIEBERS, D. L. DOWE. *Computer models solving intelligence test problems: Progress and implications*, in "Artificial Intelligence", 2016, vol. 230, pp. 74–107

[33] B. M. LAKE, T. D. ULLMAN, J. B. TENENBAUM, S. J. GERSHMAN. *Building machines that learn and think like people*, in "arXiv preprint arXiv:1604.00289", 2016

[34] C. LU, X. TANG. *Surpassing human-level face verification performance on LFW with GaussianFace*, in "arXiv preprint arXiv:1404.3840", 2014

[35] S. T. MUELLER. *A partial implementation of the BICA cognitive decathlon using the Psychology Experiment Building Language (PEBL)*, in "International Journal of Machine Consciousness", 2010, vol. 2, n^{o} 02, pp. 273–288

[36] D. SILVER, A. HUANG, C. J. MADDISON, A. GUEZ, L. SIFRE, G. VAN DEN DRIESSCHE, J. SCHRITTWIESER, I. ANTONOGLOU, V. PANNEERSHELVAM, M. LANCTOT, S. DIELEMAN, D. GREWE, J. NHAM, N. KALCHBRENNER, I. SUTSKEVER, T. LILLICRAP, M. LEACH, K. KAVUKCUOGLU, T. GRAEPEL, D. HASSABIS. *Mastering the game of Go with deep neural networks and tree search*, in "Nature", 2016, vol. 529, n^{o} 7587, pp. 484–489

[37] I. SUTSKEVER, O. VINYALS, Q. V. LE. *Sequence to sequence learning with neural networks*, in "Advances in neural information processing systems", 2014, pp. 3104–3112

[38] A. M. TURING. *Computing machinery and intelligence*, in "Mind", 1950, vol. 59, n^{o} 236, pp. 433–460

[39] W. XIONG, J. DROPPO, X. HUANG, F. SEIDE, M. SELTZER, A. STOLCKE, D. YU, G. ZWEIG. *Achieving human parity in conversational speech recognition*, in "arXiv preprint arXiv:1610.05256", 2016