RESEARCH CENTRE

Inria Paris Center

2022
ACTIVITY REPORT

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)

IN COLLABORATION WITH: Laboratoire de sciences cognitives et psycholinguistique

DOMAIN

Perception, Cognition and Interaction

THEME

Language, Speech and Audio

Contents

Team COML

Creation of the Team: 2017 May 04

Keywords

Computer sciences and digital sciences

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 members, visitors, external collaborators

Research Scientists

- Emmanuel Dupoux [Team leader, EHESS, Senior Researcher, HDR]
- Justine Cassell [INRIA, Senior Researcher, until Sep 2022]

Post-Doctoral Fellows

- Emer Gilmartin [TRINITY COLLEGE, from Aug 2022]
- William Havard [ENS, from Apr 2022 until Sep 2022]

PhD Students

- Alafate Abulimiti [INRIA]
- Robin Algayres [ENS PARIS]
- Maureen De Seyssel [ENS PARIS]
- Marvin Lavechin [ENS PARIS]
- Biswesh Mohapatra [INRIA, from Oct 2022]
- Mathieu Rita [INRIA]

Technical Staff

- Julie Bonnaire [INRIA, Engineer]
- Xuan-Nga Cao [EHESS, Engineer]
- Nicolas Hamilakis [ENS, Engineer]
- Manel Khentout [ENS, Engineer]
- Marianne Metais [ENS, Engineer]
- Maxime Poli [ENS, Engineer, from Oct 2022]
- Robin San Roman [ENS, Engineer, until Apr 2022]
- Valentin Taillandier [ENS, Engineer, until Jul 2022]
- Hadrien Titeux [ENS PARIS, Engineer]
- Catherine Urban [ENS, Engineer, until Feb 2022]
- Gwendal Virlet [ENS, Engineer, until Sep 2022]
- Sabrina Zermani [ENS, Engineer, from Mar 2022]

Interns and Apprentices

- Sarah Bahli [INRIA, from Mar 2022]
- Pablo Jose Diego Simon [INRIA, Intern, from Sep 2022]
- Gustav Grimberg [DI-ENS, from May 2022]
- Jing Liu [ENS, Intern, from Feb 2022]
- Noé Malais [ENS, Intern, from Apr 2022 until Aug 2022]
- Ewen Michel [INRIA]
- Angelo Ortiz Tandazo [ENS, Intern, from Oct 2022]
- Diana Persico [INRIA, from Mar 2022 until Aug 2022]
- Maxime Poli [INRIA Paris, Intern, from Apr 2022 until Sep 2022]
- Nina Ryszfeld [INRIA, from Nov 2022]
- Arthur Thomas [ENS, Intern, from Jun 2022 until Aug 2022]

Administrative Assistant

- Meriem Guemair [INRIA]

External Collaborators

- Ewan Dunbar [Université de Toronto]
- Abdellah Fourtassi [Université d'Aix-Marseille]
- Paul Michel [ENS, until Aug 2022]
- Juliette Millet [ENS, until May 2022]
- Tu Anh Nguyen [ENS]
- Rachid Riad [ENS, from Jun 2022]
- Sho Tsuji [UNIV TOKYO, from Apr 2022]

2 Overall objectives

Brain-inspired machine learning algorithms combined with massive datasets have recently reached spectacular results, equaling or beating humans on specific high level tasks (e.g. the game of go, poker, diplomacy) and achieving uncanny abilities in generating images or text. However, there are still a lot of domains in which even humans infants outperform machines: data-limited 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 au,tonomous and data efficient 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.

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., [32, 36, 41, 31, 38]). 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) [2]; 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) [34].

The general aim of CoML, following the roadmap described in [2], 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, 5, 7], 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 [39].

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 [40] 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* [33] , i.e., functional components within a more global cognitive architecture [37]. 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 [6], [35].

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 Social and environmental responsibility

5.1 Footprint of research activities

The footprint of the CoML team due to travel has increased since the covid time, but did not reach the pre-covid phase, since many conferences were attended remotely. The compute footprint was that of our 4 GPU cluster, used on average 30% of the time, to which we should add the compute of our accounts at Jean Zay (we could not access the data at the time of the report).

5.2 Impact of research results

Our fundamental work in unsupervised learning algorithms are very early stage and have up to now no known environmental/societal application. Our applicative work is dedicated to develop spoken language annotation and analysis tools for researchers, which should help conduct research in clinical and developmental areas (Health and Well Being, and early Education).

6 New software and platforms

6.1 New software

6.1.1 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

News of the Year: New processors for Vocal Tract Length Normalization and pitch extraction.

URL: https://docs.cognitive-ml.fr/shennong

Contact: Mathieu Bernard

6.1.2 phonemizer

Keyword: Text

Functional Description: * Conversion of a text into its phonemic representation * Wrapper on speech synthesis programs espeak and festival

News of the Year: Support for SAMPA phonetic alphabet with the new espeak-sampa backend. A lot of improvments and bug fixes.

URL: https://github.com/bootphon/phonemizer

Contact: Mathieu Bernard

6.1.3 TDE

Name: Term Discovery Evaluation

Keywords: NLP, Speech recognition, Speech

Scientific Description: This toolbox allows the user to judge of the quality of a word discovery algorithm. It evaluates the algorithms on these criteria : - Boundary : efficiency of the algorithm to found the actual boundaries of the words - Group : efficiency of the algorithm to group similar words - Token/Type: efficiency of the algorithm to find all words from the corpus (types), and to find all occurences (token) of these words. - NED : Mean of the edit distance across all the word pairs found by the algorithm - Coverage : efficiency of the algorithm to find every discoverable phone in the corpus

Functional Description: Toolbox to evaluate algorithms that segment speech into words. It allows the user to evaluate the efficiency of algorithms to segment speech into words, and create clusters of similar words.

News of the Year: Complete rewrite (optimization and bugfixes)

URL: https://github.com/bootphon/tdev2

Contact: Emmanuel Dupoux

6.1.4 wordseg

Name: wordseg

Keywords: Segmentation, Text, NLP

Functional Description: * Provides a collection of tools for text based word segmentation. * Covers the whole segmentation pipeline: data preprocessing, algorithms, evaluation and descriptive statistics. * Implements 6 segmentation algorithms and a baseline * Available as a Python library and command-line tools

News of the Year: New functionalities for cross-validation.

URL: https://wordseg.readthedocs.io

Contact: Mathieu Bernard

Partner: ENS Paris

6.1.5 abkhazia

Keywords: Speech recognition, Speech-text alignment

Functional Description: The Abkhazia sofware makes it easy to obtain simple baselines for supervised ASR (using Kaldi) and ABX tasks (using ABXpy) on the large corpora of speech recordings typically used in speech engineering, linguistics or cognitive science research.

URL: https://github.com/bootphon/abkhazia

Contact: Emmanuel Dupoux

6.1.6 ABXpy

Keywords: Evaluation, Speech recognition, Machine learning

Functional Description: The ABX package gives a performance score to speech recognition systems by measuring their capacity to discriminate linguistic contrasts (accents, phonemes, speakers, etc...)

URL: https://github.com/bootphon/ABXpy

Contact: Emmanuel Dupoux

6.1.7 abnet3

Keywords: Artificial intelligence, Speech processing, Deep learning, Unsupervised learning

Functional Description: Siamese network for unsupervised speech representation learning

URL: https://github.com/bootphon/abnet3

Contact: Emmanuel Dupoux

6.1.8 h5features

Keyword: File format

Functional Description: The h5features python package provides easy to use and efficient storage of large features data on the HDF5 binary file format.

URL: https://github.com/bootphon/h5features

Contact: Emmanuel Dupoux

6.1.9 intphys

Name: 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..

URL: http://www.intphys.com

Contact: Mathieu Bernard

6.1.10 Seshat

Name: 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.

URL: https://github.com/bootphon/seshat

Contact: Hadrien Titeux

Partner: ENS Paris

6.1.11 pyGammaAgreement

Name: pyGammaAgreement

Keywords: Reliability, Measures

Functional Description: Python library for measuring inter and intra annotator reliability for annotation sequences

URL: https://github.com/bootphon/pygamma-agreement

Contact: Emmanuel Dupoux

7 New results

Participants: Emmanuel Dupoux, Justine Cassel.

7.1 Unsupervised learning

Humans learn to speak and to perceive the world in a largely self-supervised fashion from raw sensory data. Yet, most of machine learning is still devoted to partly supervised algorithms that rely on human labels, or work on textual inputs, which is not how humans learn language. We have used infants as sources of inspiration for developing novel machine learning algorithms that learn from raw audio inputs without any text. This opens up a new area of research that we have called ’textless NLP’ (see the Generative Spoken Language Model introduced last year in [4]).

- Some of our past work in spoken language modeling shows the possibility of learning a language unsupervisedly from raw audio without any text labels. The approach relies first on transforming the audio into a sequence of discrete units (or pseudo-text) and then training a language model directly on such pseudo-text. Is such a discrete bottleneck necessary, potentially introducing irreversible errors in the encoding of the speech signal, or could we learn a language model without discrete units at all? In [13], we study the role of discrete versus continuous representations in spoken language modeling. We show that discretization is indeed essential for good results in spoken language modeling. We show that discretization removes linguistically irrelevant information from the continuous features, helping to improve language modeling performances. On the basis of this study, we train a language model on the discrete units of the HuBERT features, reaching new state-of-the-art results in the lexical, syntactic and semantic metrics of the Zero Resource Speech Challenge 2021 (Track 1 - Speech Only).

- Speech pre-training has primarily demonstrated efficacy on classification tasks, while its capability of generating novel speech, similar to how GPT-2 can generate coherent paragraphs, has barely been explored. Generative Spoken Language Modeling (GSLM) [4] is the only prior work addressing the generative aspects of speech pretraining, which replaces text with discovered phone-like units for language modeling and shows the ability to generate meaningful novel sentences. Unfortunately, despite eliminating the need of text, the units used in GSLM discard most of the prosodic information. Hence, GSLM fails to leverage prosody for better comprehension, and does not generate expressive speech. In [17], we present a prosody-aware generative spoken language model (pGSLM). It is composed of a multi-stream transformer language model (MS-TLM) of speech, represented as discovered unit and prosodic feature streams, and an adapted HiFi-GAN model converting MSTLM outputs to waveforms. We devise a series of metrics for prosody modeling and generation, and re-use metrics from GSLM for content modeling. Experimental results show that the pGSLM can utilize prosody to improve both prosody and content modeling, and also generate natural, meaningful, and coherent speech given a spoken prompt. Samples are available under the following link

- Textless spoken language processing research aims to extend the applicability of standard NLP toolset onto spoken language and languages with few or no textual resources. In [16], we introduce textless-lib, a PyTorch-based library aimed to facilitate research in this research area. We describe the building blocks that the library provides and demonstrate its usability by discuss three different use-case examples: (i) speaker probing, (ii) speech resynthesis and compression, and (iii) speech continuation. We believe that textless-lib substantially simplifies research the textless setting and will be handful not only for speech researchers but also for the NLP community at large. The code, documentation, and pre-trained models are available at this link.

- Finding word boundaries in continuous speech is challenging as there is little or no equivalent of a ‘space’ delimiter between words. Popular Bayesian non-parametric models for text segmentation (Goldwater et al., 2006, 2009) use a Dirichlet process to jointly segment sentences and build a lexicon of word types. In [8], we introduce DP-Parse, which uses similar principles but only relies on an instance lexicon of word tokens, avoiding the clustering errors that arise with a lexicon of word types. On the Zero Resource Speech Benchmark 2017, our model sets a new speech segmentation state-of-theart in 5 languages. The algorithm monotonically improves with better input representations, achieving yet higher scores when fed with weakly supervised inputs. Despite lacking a type lexicon, DP-Parse can be pipelined to a language model and learn semantic and syntactic representations as assessed by a new spoken word embedding benchmark.

- In [15], we introduce a simple neural encoder architecture that can be trained using an unsupervised contrastive learning objective which gets its positive samples from data-augmented k-Nearest Neighbors search. We show that when built on top of recent self-supervised audio representations, this method can be applied iteratively and yield competitive SSE as evaluated on two tasks: query-by-example of random sequences of speech, and spoken term discovery. On both tasks our method pushes the state-of-the-art by a significant margin across 5 different languages. Finally, we establish a benchmark on a query-by-example task on the LibriSpeech dataset to monitor future improvements in the field.

- In [18], we explore a possible application of textless NLP: Speech emotion conversion is the task of modifying the perceived emotion of a speech utterance while preserving the lexical content and speaker identity; it is traditionally done at the signal level with not very convincing results. We cast emotion conversion as a textless spoken language translation task: speech signals are decomposed into discrete learned representations, consisting of phonetic-content units, prosodic features, speaker, and emotion. We then modify the speech content by translating the phonetic content units to a target emotion, and then predict the prosodic features based on these units. Finally, the speech waveform is generated by feeding the predicted representations into a neural vocoder. Such a paradigm allows us to go beyond spectral and parametric changes of the signal, and model non-verbal vocalizations, such as laughter insertion, yawning removal, etc. We demonstrate objectively and subjectively that the proposed method is vastly superior to current approaches and even beats text-based systems in terms of perceived emotion and audio quality. We rigorously evaluate all components of such a complex system and conclude with an extensive model analysis and ablation study to better emphasize the architectural choices, strengths and weaknesses of the proposed method. Samples are available under the following link.

7.2 Datasets and Benchmarks

Self-supervised learning is a relatively new field of research. The CoML team contributes to the research by building benchmarks and organizing challenges to enable comparison between systems on a single set of metrics and cumulative progress across laboratories. The specificity of our approach is that we base our metrics on human psycholinguistics and psychophysics, enablig direct human - machine comparisons.

- End-to-end spoken language understanding (SLU) predicts intent directly from audio using a single model. It promises to improve the performance of assistant systems by leveraging acoustic information lost in the intermediate textual representation and preventing cascading errors from Automatic Speech Recognition (ASR). Further, having one unified model has efficiency advantages when deploying assistant systems on-device. Unfortunately, the limited number of public audio datasets with semantic parse labels hinders the research progress in this area. In [27], we release the Spoken task-oriented semantic parsing (STOP) dataset, the largest and most complex SLU dataset to be publicly available. Additionally, we define low-resource splits to establish a benchmark for improving SLU when limited labeled data is available. Furthermore, in addition to the human-recorded audio, we are releasing a TTS-generated version to benchmark the performance for low-resource domain adaptation of end-to-end SLU systems.

- Recent progress in self-supervised or unsupervised machine learning has opened the possibility of building a full speech processing system from raw audio without using any textual representations or expert labels such as phonemes, dictionaries or parse trees. The contribution of the Zero Resource Speech Challenge series since 2015 has been to break down this long-term objective into four well-defined tasks—Acoustic Unit Discovery, Spoken Term Discovery, Discrete Resynthesis, and Spoken Language Modeling—and introduce associated metrics and benchmarks enabling model comparison and cumulative progress. In [10], we present an overview of the six editions of this challenge series since 2015, discuss the lessons learned, and outline the areas which need more work or give puzzling results.

7.3 Language emergence in communicative agents

In this research topic, which was the focus of Rahma Chaabouni's PhD thesis [1], which was taken up by the MSR funded PhD of Mathieu Rita, we study the inductive biases of neural systems by presenting them with few or no data.

- Lewis signaling games are a class of simple communication games for simulating the emergence of language. In these games, two agents must agree on a communication protocol in order to solve a cooperative task. Previous work has shown that agents trained to play this game with reinforcement learning tend to develop languages that display undesirable properties from a linguistic point of view (lack of generalization, lack of compositionality, etc). In [23], we aim to provide better understanding of this phenomenon by analytically studying the learning problem in Lewis games. As a core contribution, we demonstrate that the standard objective in Lewis games can be decomposed in two components: a co-adaptation loss and an information loss. This decomposition enables us to surface two potential sources of overfitting, which we show may undermine the emergence of a structured communication protocol. In particular, when we control for overfitting on the co-adaptation loss, we recover desired properties in the emergent languages: they are more compositional and generalize better

- Populations have often been perceived as a structuring component for language to emerge and evolve: the larger the population, the more structured the language. While this observation is widespread in the sociolinguistic literature, it has not been consistently reproduced in computer simulations with neural agents. In [22], we thus aim to clarify this apparent contradiction. We explore emergent language properties by varying agent population size in the speaker-listener Lewis Game. After reproducing the experimental difference, we challenge the simulation assumption that the agent community is homogeneous. We then investigate how speaker-listener asymmetry alters language structure through the analysis a potential diversity factor: learning speed. From then, we leverage this observation to control population heterogeneity without introducing confounding factors. We finally show that introducing such training speed heterogeneities naturally sort out the initial contradiction: larger simulated communities start developing more stable and structured languages.

7.4 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/interpretation methods based on psycholinguistic/linguistic/neuroscience criteria, and deploy them for systematic comparison of systems.

- Work done in recent years has shown the usefulness of using automatic methods for the study of linguistic typology. However, the majority of proposed approaches come from natural language processing and require expert knowledge to predict typological information for new languages. An alternative would be to use speech-based methods that do not need extensive linguistic annotations, but considerably less work has been done in this direction. In [26] we aim to reduce this gap, by investigating a promising speech representation, i-vectors, which by capturing suprasegmental features of language, can be used for the automatic characterization of languages. Employing data from 24 languages, covering several linguistic families, we computed the i-vectors corresponding to each sentence and we represented the languages by their centroid i-vector. Analyzing the distance between the language centroids and phonological, inventory and syntactic distances between the same languages, we observed a significant correlation between the i-vector distance and the syntactic distance. Then, we explored in more detailed a number of syntactic features and we proposed a method for predicting the value of the most promising feature, based on the i-vector information. The obtained results, an 87% classification accuracy, are encouraging and we envision to extend this method further.

- Previous work has shown that Contrastive Predictive Coding (CPC) can discover representations that encode phonetic information. In [24], we ask what other types of information are present in CPC speech representations. We focus on three categories: phone class, gender and language, and compare monolingual and bilingual models. Using qualitative and quantitative tools, we find that both gender and phone class information are present in both types of models. Language information, however, is very salient in the bilingual model only, suggesting CPC models learn to discriminate languages when trained on multiple languages. Some language information can also be retrieved from monolingual models, but it is more diffused across all features. These patterns hold when analyses are carried on the discrete units from a downstream clustering model. However, although there is no effect of the number of target clusters on phone class and language information, more gender information is encoded with more clusters. Finally, we find that there is some cost to being exposed to two languages on a downstream phoneme discrimination task.

7.5 Simulation of language learning in infants

Supervised learning algorithms provide very interesting quantitative models of the early phases of language acquisition. When fed with realistic input, they can generate predictions that can be compared with available developmental behavioral data.

- Infants come to learn several hundreds of word forms by two years of age, and it is possible this involves carving these forms out from continuous speech. It has been proposed that the task is facilitated by the presence of prosodic boundaries. In [12], we revisit this claim by running computational models of word segmentation, with and without prosodic information, on a corpus of infant-directed speech. We use five cognitively-based algorithms, which vary in whether they employ a sub-lexical or a lexical segmentation strategy and whether they are simple heuristics or embody an ideal learner. Results show that providing expert-annotated prosodic breaks does not uniformly help all segmentation models. The sub-lexical algorithms, which perform more poorly, benefit most, while the lexical ones show a very small gain. Moreover, when prosodic information is derived automatically from the acoustic cues infants are known to be sensitive to, errors in the detection of the boundaries lead to smaller positive effects, and even negative ones for some algorithms. This shows that even though infants could potentially use prosodic breaks, it does not necessarily follow that they should incorporate prosody into their segmentation strategies, when confronted with realistic signals.
- In [30], we argue that language use in everyday life can be studied using lightweight, wearable recorders that collect long-form recordings—that is, audio (including speech) over whole days. The hardware and software underlying this technique are increasingly accessible and inexpensive, and these data are revolutionizing the language acquisition field. We first place this technique into the broader context of the current ways of studying both the input being received by children and children's own language production, laying out the main advantages and drawbacks of long-form recordings. We then go on to argue that a unique advantage of long-form recordings is that they can fuel realistic models of early language acquisition that use speech to represent children's input and/or to establish production benchmarks. To enable the field to make the most of this unique empirical and conceptual contribution, we outline what this reverse engineering approach from long-form recordings entails, why it is useful, and how to evaluate success.

7.6 Quantitative studies of human learning and processing

In this section, we focus on the use of machine learning algorithms of speech and language processing to study speech processing in humans. We do this in two ways, First by using the systems as models of humans speech processing and attempting to reproduce known behavioral or brain results (the fist two papers). Second, by providing tools to quantify speech processing in the case of patients with neurodegenerative diseases. This last topic (the last three papers) reflect the work of Rachid Riad who defended his PhD in 2022 [28].

- According to the Language Familiarity Effect (LFE), people are better at discriminating between speakers of their native language. Although this cognitive effect was largely studied in the literature, experiments have only been conducted on a limited number of language pairs and their results only show the presence of the effect without yielding a gradual measure that may vary across language pairs. In this work, we show that the computational model of LFE introduced by Thorburn, Feldman, and Schatz (2019) can address these two limitations. In [25], we attest to this model's capacity to obtain a gradual measure of the LFE by replicating behavioural findings on native and accented speech. We then evaluate LFE on a large number of language pairs, including many which have never been tested on humans. We show that the effect is replicated across a wide array of languages, providing further evidence of its universality. Building on the gradual measure of LFE, we also show that languages belonging to the same family yield smaller scores, supporting the idea of an effect of language distance on LFE.

- Several deep neural networks have recently been shown to generate activations similar to those of the brain in response to the same input. These algorithms, however, remain largely implausible: they require (1) extraordinarily large amounts of data, (2) unobtainable supervised labels, (3) textual rather than raw sensory input, and / or (4) implausibly large memory (e.g. thousands of contextual words). These elements highlight the need to identify algorithms that, under these limitations, would suffice to account for both behavioral and brain responses. Focusing on speech processing, in [19], we hypothesize that self supervised algorithms trained on the raw waveform constitute a promising candidate. Specifically, we compare a recent self-supervised model, wav2vec 2.0, to the brain activity of 412 English, French, and Mandarin individuals recorded with functional Magnetic Resonance Imaging (fMRI), while they listened to approximately one hour of audio books. First, we show that this algorithm learns brain-like representations with as little as 600 hours of unlabelled speech – a quantity comparable to what infants can be exposed to during language acquisition. Second, its functional hierarchy aligns with the cortical hierarchy of speech processing. Third, different training regimes reveal a functional specialization akin to the cortex: wav2vec 2.0 learns sound-generic, speech-specific and language-specific representations similar to those of the prefrontal and temporal cortices. Fourth, we confirm the similarity of this specialization with the behavior of 386 additional participants. These elements, resulting from the largest neuroimaging benchmark to date, show how self supervised learning can account for a rich organization of speech processing in the brain, and thus delineate a path to identify the laws of language acquisition which shape the human brain.

- Conversations between a clinician and a patient, in natural conditions, are valuable sources of information for medical follow-up. The automatic analysis of these dialogues could help extract new language markers and speed up the clinicians' reports. Yet, it is not clear which model is the most efficient to detect and identify the speaker turns, especially for individuals with speech disorders. In [21], we proposed a split of the data that allows conducting a comparative evaluation of different diarization methods. We designed and trained end-to-end neural network architectures to directly tackle this task from the raw signal and evaluate each approach under the same metric. We also studied the effect of fine-tuning models to find the best performance. Experimental results are reported on naturalistic clinical conversations between Psychologists and Interviewees, at different stages of Huntington's disease, displaying a large panel of speech disorders. We found out that our best end-to-end model achieved 19.5% IER on the test set, compared to 23.6% achieved by the finetuning of the X-vector architecture. Finally, we observed that we could extract clinical markers directly from the automatic systems, highlighting the clinical relevance of our methods.

- Using brief samples of speech recordings, in [14], we aimed at predicting, through machine learning, the clinical performance in Huntington's Disease (HD), an inherited Neurodegenerative disease (NDD). Methods We collected and analyzed 126 samples of audio recordings of both forward and backward counting from 103 Huntington's disease gene carriers [87 manifest and 16 premanifest; mean age 50.6 (SD 11.2), range (27–88) years] from three multicenter prospective studies in France and Belgium (MIG-HD (ClinicalTrials.gov NCT00190450); BIO-HD (ClinicalTrials.gov NCT00190450) and Repair-HD (ClinicalTrials.gov NCT00190450). We pre-registered all of our methods before running any analyses, in order to avoid infated results. We automatically extracted 60 speech features from blindly annotated samples. We used machine learning models to combine multiple speech features in order to make predictions at individual levels of the clinical markers. We trained machine learning models on 86% of the samples, the remaining 14% constituted the independent test set. We combined speech features with demographics variables (age, sex, CAG repeats, and burden score) to predict cognitive, motor, and functional scores of the Unifed Huntington's disease rating scale. We provided correlation between speech variables and striatal volumes. Results Speech features combined with demographics allowed the prediction of the individual cognitive, motor, and functional scores with a relative error from 12.7 to 20.0% which is better than predictions using demographics and genetic information. Both mean and standard deviation of pause durations during backward recitation and clinical scores correlated with striatal atrophy (Spearman 0.6 and 0.5–0.6, respectively). Interpretation Brief and examiner-free speech recording and analysis may become in the future an efcient method for remote evaluation of the individual condition in HD and likely in other NDD.

- Patients with Huntington's disease suffer from disturbances in the perception of emotions; they do not correctly read the body, vocal and facial expressions of others. With regard to the expression of emotions, it has been shown that they are impaired in expressing emotions through face but up until now, little research has been conducted about their ability to express emotions through spoken language. In [11], to better understand emotion production in both voice and language in Huntington's Disease (HD), we tested 115 individuals: 68 patients (HD), 22 participants carrying the mutant HD gene without any motor symptoms (pre-manifest HD), and 25 controls in a single-centre prospective observational follow-up study. Participants were recorded in interviews in which they were asked to recall sad, angry, happy, and neutral stories. Emotion expression through voice and language was investigated by comparing the identifiability of emotions expressed by controls, preHD and HD patients in these interviews. To assess separately vocal and linguistic expression of emotions in a blind design, we used machine learning models instead of a human jury performing a forced-choice recognition test. Results from this study showed that patients with HD had difficulty expressing emotions through both voice and language compared to preHD participants and controls, who behaved similarly and above chance. In addition, we did not find any differences in expression of emotions between preHD and healthy controls. We further validated our newly proposed methodology with a human jury on the speech produced by the controls. These results are consistent with the hypothesis that emotional deficits in HD are caused by impaired sensori-motor representations of emotions, in line with embodied cognition theories. This study also shows how machine learning models can be leveraged to assess emotion expression in a blind and reproducible way.

7.7 Interactive AI

Interactive AI focuses on the ways in which we ensure that AI systems conduct conversations with people in the most natural and effective ways. We accomplish this goal by conducting research on human-human conversational interaction, and then integrating the results into AI systems. Two approaches have characterized our work in 2022: (1) examining linguistic phenomena (including words, but also nonverbal behavior such as head nods, eye gaze, or hand gestures and speech phenomena such as prosody) that make dialogue effective, and successful, and (2) examining the ways in which these same linguistic and nonverbal phenomena allow people to knit social bonds with one another. (1) The past couple of years have seen the rapid introduction and growth of large language models (LLM trained on huge corpora of dialogues (such as BlenderBot, LaMDA, Dialo-GPT and ChatGPT). In this context it becomes increasingly important to define what makes natural dialogue or conversation both effective and satisfying, and to ensure that these characteristics are automatically detected by these models, and also generated by such large dialogue models. This work has become a primary focus of my work and that of my students. (2) LLM are have until today been incapable of generating interactional phenomena. This has changed with ChatGPT. For example, if one asks it "please tell somebody to wash their clothes using a hedge" it gives back "you might want to wash your clothes" That's excellent progress, but even ChatGPT does not know *when* to generate hedges. In people, hedges are used early in a relationship or in situations where the social bonds between speaker and listener are weak. We are therefore examining the conditions under which social bonds are formed between people in a conversation and *how* these bonds are formed. We accomplish this by looking at data from 3 sources: analyses of linguistic and nonverbal data, hyperscanning data that measures the synchrony of brain waves between two individuals, and performance on collaborative tasks.

- Hedges play an important role in the management of conversational interaction. In peer tutoring, they are notably used by tutors in dyads (pairs of interlocutors) experiencing low rapport to tone down the impact of instructions and negative feedback. Pursuing the objective of building a tutoring agent that manages rapport with students in order to improve learning, in [20] , we used a multimodal peer-tutoring dataset to construct a computational framework for identifying hedges. We compared approaches relying on pre-trained resources with others that integrate insights from the social science literature. Our best performance involved a hybrid approach that outperforms the existing baseline while being easier to interpret. We employ a model explainability tool to explore the features that characterize hedges in peer-tutoring conversations, and we identify some novel features, and the benefits of such a hybrid model approach.

- Socio-conversational systems are dialogue systems, including what are sometimes referred to as chatbots, vocal assistants, social robots, and embodied conversational agents, that are capable of interacting with humans in a way that treats both the specifically social nature of the interaction and the content of a task. In [9], we 1) uncover some places where the compartmentalized nature of research conducted around socio-conversational systems creates problems for the field as a whole, and 2) propose a way to overcome this compartmentalization and thus strengthen the capabilities of socio-conversational systems by defining common challenges. Specifically, we examine research carried out by the signal processing, natural language processing and dialogue, machine deep learning, social/affective computing and social sciences communities. We focus on three major challenges for the development of effective socio-conversational systems, and describe ways to tackle them.

- Children's interaction with peers is enormously influential in their cognitive, social, and emotional development. This fact led to the development of child-like embodied conversational agents called virtual peers. In this chapter, I focus on those virtual peers—the kinds of ECAs and robots where the computer takes on the role of a peer, often communicating with ageappropriate language, and even looking like a child of the same age as the young person interacting with it. For the most part, the application of contemporary virtual peers is learning, and so in this chapter I narrow the focus to learning (as opposed to interactions to simply pass the time without other goals, for example) among students up through university, but we broaden the discussion beyond the classic school curriculum to informal learning outside the classroom. I also broaden the focus beyond what are sometimes called "core literacies"—reading, writing, and arithmetic—to include the learning of socio-emotional skills such as curiosity and establishing social bonds.

- Curiosity is a vital metacognitive skill in educational contexts, leading to creativity, and a love of learning. And while many school systems increasingly undercut curiosity by teaching to the test, teachers are increasingly interested in how to evoke curiosity in their students to prepare them for a world in which lifelong learning and reskilling will be more and more important. One aspect of curiosity that has received little attention, however, is the role of peers in eliciting curiosity. In [29], we present what we believe to be the first theoretical framework that articulates an integrated socio-cognitive account of curiosity that ties observable behaviors in peers to underlying curiosity states. We make a bipartite distinction between individual and interpersonal functions that contribute to curiosity, and multimodal behaviors that fulfill these functions. We validate the proposed framework by leveraging a longitudinal latent variable modeling approach. Findings confirm a positive predictive relationship between the latent variables of individual and interpersonal functions and curiosity, with the interpersonal functions exercising a comparatively stronger influence. Prominent behavioral realizations of these functions are also discovered in a data-driven manner. We instantiate the proposed theoretical framework in a set of strategies and tactics that can be incorporated into learning technologies to indicate, evoke, and scaffold curiosity. This work is a step towards designing learning technologies that can recognize and evoke moment-by-moment curiosity during learning in social contexts and towards a more complete multimodal learning analytics. The underlying rationale is applicable more generally for developing computer support for other metacognitive and socio-emotional skills.

8 Bilateral contracts and grants with industry

Participants: Emmanuel Dupoux.

- **Facebook AI Research Grant** (2022, PI: E. Dupoux, 350K€) - Unrestricted Gift - The aim is to help the development of machine learning tools geared towards the psycholinguistic research community.

9 Partnerships and cooperations

Participants: Emmanuel Dupoux.

9.1 Regional initiatives

- **SESAME Echolalia.** (2018-2021; coordinating PI : E. Dupoux; 400K€) - Development of an open source speech and video secure annotation platform.

10 Dissemination

Participants: Emmanuel Dupoux, Justine Cassel.

10.1 Promoting scientific activities

10.1.1 Invited talks

E. Dupoux gave a keynote lecture at LREC, June 2022, Marseille on Textless NLP.

10.2 Teaching - Supervision - Juries

10.2.1 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
- Doctorat : E. Dupoux, "Computational models of cognitive development", 32 h, Séminaire EHESS, Paris France

J. Cassell is professor of language technologies and human-computer interaction (ENS)

- Master: J. Cassell "Conversation among People and Bots", fall semester, M2 (Cogmaster), ENS-EHESS-UP
- Guest lectures/seminars in classes: J. Cassell: EPFL, University of Pennsylvania, UE Cognition Sociale de la Sorbonne, Seminaire Doctoral Littérature et Culture d'Enfance at ENS-Afreloce, Ethics & Societal Impact of AI at MBZUAI (MBZ University of AI in Abu Dhabi), Sociologie du Numérique.

11 Scientific production

11.1 Major publications

[1] R. Chaabouni. 'Emerging linguistic universals in communicating neural network agents'. Université Paris sciences et lettres, 17th Mar. 2021. URL: https://hal.inria.fr/tel-03536320.

[2] E. Dupoux. 'Cognitive Science in the era of Artificial Intelligence: A roadmap for reverse-engineering the infant language-learner'. In: *Cognition* (2018).

[3] A. Fourtassi, T. Schatz, B. Varadarajan and 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*. Vol. 2. ACL. Baltimore, Maryland: Association for Computational Linguistics, 2014, pp. 1–6. DOI: 10.3115/v1/P14-2001.

[4] K. Lakhotia, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baevski, A. Mohamed and E. Dupoux. 'On Generative Spoken Language Modeling from Raw Audio'. In: *Transactions of the Association for Computational Linguistics* (1st Feb. 2021). DOI: 10.1162/tacl_a_00430. URL: https://hal.inria.fr/hal-03329219.

[5] A. Martin, S. Peperkamp and E. Dupoux. 'Learning Phonemes with a Proto-lexicon'. In: *Cognitive Science* 37 (2013), pp. 103–124. DOI: 10.1111/j.1551-6709.2012.01267.x.

[6] T. Schatz, V. Peddinti, F. Bach, A. Jansen, H. Hynek and E. Dupoux. 'Evaluating speech features with the Minimal-Pair ABX task: Analysis of the classical MFC/PLP pipeline'. In: *INTERSPEECH-2013*. International Speech Communication Association. Lyon, France, 2013, pp. 1781–1785.

[7] R. Thiollière, E. Dunbar, G. Synnaeve, M. Versteegh and E. Dupoux. 'A Hybrid Dynamic Time Warping-Deep Neural Network Architecture for Unsupervised Acoustic Modeling'. In: *INTERSPEECH-2015*. 2015, pp. 3179–3183.

11.2 Publications of the year

International journals

[8] R. Algayres, T. Ricoul, J. Karadayi, H. Laurençon, S. Zaiem, A. Mohamed, B. Sagot and E. Dupoux. 'DP-Parse: Finding Word Boundaries from Raw Speech with an Instance Lexicon'. In: *Transactions of the Association for Computational Linguistics* 10 (19th Sept. 2022), pp. 1051–1065. DOI: 10.1162/tacl_a_00505. URL: https://hal.inria.fr/hal-03831873.

[9] C. Clavel, M. Labeau and J. Cassell. 'Socio-conversational systems: Three challenges at the crossroads of fields'. In: *Frontiers in Robotics and AI* 9 (15th Dec. 2022). DOI: 10.3389/frobt.2022.937825. URL: https://hal.inria.fr/hal-03990543.

[10] E. Dunbar, N. Hamilakis and E. Dupoux. 'Self-supervised language learning from raw audio: Lessons from the Zero Resource Speech Challenge'. In: *IEEE Journal of Selected Topics in Signal Processing* (2022). DOI: 10.1109/jstsp.2022.3206084. URL: https://hal.science/hal-03789716.

[11] C. Gallezot, R. Riad, H. Titeux, L. Lemoine, J. Montillot, A. Sliwinski, J. H. Bagnou, X. N. Cao, K. Youssov, E. Dupoux and A.-C. Bachoud Levi. 'Emotion expression through spoken language in Huntington disease'. In: *Cortex* 155 (19th July 2022), pp. 150–161. DOI: 10.1016/j.cortex.2022.05.024. URL: https://hal.science/hal-03993326.

[12] B. Ludusan, A. Cristia, R. Mazuka and E. Dupoux. 'How much does prosody help word segmentation? A simulation study on infant-directed speech'. In: *Cognition* 219 (Feb. 2022), p. 104961. DOI: 10.1016/j.cognition.2021.104961. URL: https://hal.science/hal-03991055.

[13] T. A. Nguyen, B. Sagot and E. Dupoux. 'Are discrete units necessary for Spoken Language Modeling?' In: *IEEE Journal of Selected Topics in Signal Processing* (23rd Aug. 2022). URL: https://hal.inria.fr/hal-03831707.

[14] R. Riad, M. Lunven, H. Titeux, X.-N. Cao, J. Hamet Bagnou, L. Lemoine, J. Montillot, A. Sliwinski, K. Youssov, L. Cleret de Langavant, E. Dupoux and A.-C. Bachoud-Lévi. 'Predicting clinical scores in Huntington's disease: a lightweight speech test'. In: *Journal of Neurology* 269.9 (14th May 2022), pp. 5008–5021. DOI: 10.1007/s00415-022-11148-1. URL: https://hal.inria.fr/hal-03831681.

International peer-reviewed conferences

[15] R. Algayres, A. Nabli, B. Sagot and E. Dupoux. 'Speech Sequence Embeddings using Nearest Neighbors Contrastive Learning'. In: Interspeech 2022 - 23rd INTERSPEECH Conference. Incheon, South Korea, 18th Sept. 2022. URL: https://hal.inria.fr/hal-03831888.

[16] E. Kharitonov, J. Copet, K. Lakhotia, T. A. Nguyen, P. Tomasello, A. Lee, A. Elkahky, W.-N. Hsu, A. Mohamed, E. Dupoux and Y. Adi. 'Textless-lib: a Library for Textless Spoken Language Processing'. In: NAACL 2022 - Annual Conference of the North American Chapter of the Association for Computational Linguistics. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations. Seattle, United States, 10th July 2022, pp. 1–9. URL: https://hal.inria.fr/hal-03831838.

[17] E. Kharitonov, A. Lee, A. Polyak, Y. Adi, J. Copet, K. Lakhotia, T.-A. Nguyen, M. Rivière, A. Mohamed, E. Dupoux and W.-N. Hsu. 'Text-Free Prosody-Aware Generative Spoken Language Modeling'. In: ACL 2022 - Association for Computational Linguistics. Vol. 1: Long Papers. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin, Ireland: MIT Press, 22nd May 2022, pp. 8666–8681. URL: https://hal.inria.fr/hal-03831818.

[18] F. Kreuk, A. Polyak, J. Copet, E. Kharitonov, T.-A. Nguyen, M. Rivière, W.-N. Hsu, A. Mohamed, E. Dupoux and Y. Adi. 'Textless Speech Emotion Conversion using Discrete & Decomposed Representations'. In: *Proceedings of EMNLP*. EMNLP 2022. Abu Dhabi (online), United Arab Emirates, 7th Dec. 2022. URL: https://hal.inria.fr/hal-03831801.

[19] J. Millet, C. Caucheteux, P. Orhan, Y. Boubenec, A. Gramfort, C. C. Pallier, E. Dunbar and J.-R. King. 'Toward a realistic model of speech processing in the brain with self-supervised learning'. In: NeurIPS 2022 - 36th Conference on Neural Information Processing Systems. New Orleans, United States, 28th Nov. 2022. URL: https://hal.science/hal-03808200.

[20] Y. Raphalen, C. Clavel and J. Cassell. '"You might think about slightly revising the title": Identifying Hedges in Peer-tutoring Interactions'. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers, "You might think about slightly revising the title": Identifying Hedges in Peer-tutoring Interactions.* ACL 2022 - 60th Annual Meeting of the Association for Computational Linguistics. Vol. 1. Dublin, Ireland: Association for Computational Linguistics, 1st May 2022, pp. 2160–2174. DOI: 10.18653/v1/2022.acl-long.153. URL: https://hal.inria.fr/hal-03990509.

[21] R. Riad, H. Titeux, X. N. Cao, E. Dupoux, L. Lemoine, J. Montillot, A. Sliwinski, J. H. Bagnou and A.-C. Bachoud-Lévi. 'A comparison study on patient-psychologist voice diarization'. In: SLPAT 2022 - 9th Workshop on Speech and Language Processing for Assistive Technologies. Dublin, Ireland, 27th May 2022. URL: https://hal.inria.fr/hal-03831674.

[22] M. Rita, F. Strub, J.-B. Grill, O. Pietquin and E. Dupoux. 'On the role of population heterogeneity in emergent communication'. In: ICLP 2022 - Tenth International Conference on Learning Representations. Los Angeles, United States, 25th Apr. 2022. URL: https://hal.inria.fr/hal-03830500.

[23] M. Rita, C. Tallec, P. Michel, J.-B. Grill, O. Pietquin, E. Dupoux and F. Strub. 'Emergent Communication: Generalization and Overfitting in Lewis Games'. In: NeurIPS 2022 - Thirty-sixth Conference on Neural Information Processing Systems. Nouvelle-Orléans, United States, 28th Nov. 2022. URL: https://hal.inria.fr/hal-03831908.

[24] M. de Seyssel, M. Lavechin, Y. Adi, E. Dupoux and G. Wisniewski. 'Probing phoneme, language and speaker information in unsupervised speech representations'. In: Interspeech 2022 - 23rd INTERSPEECH Conference. Incheon, South Korea, 18th Sept. 2022. URL: https://hal.inria.fr/hal-03830470.

[25] M. de Seyssel, G. Wisniewski and E. Dupoux. 'Is the Language Familiarity Effect gradual ? A computational modelling approach'. In: CogSci 2022 - 44th Annual Meeting of the Cognitive Science Society. Toronto, Canada, 27th July 2022. URL: https://hal.inria.fr/hal-03830461.

[26] M. D. Seyssel, G. Wisniewski, E. Dupoux and B. Ludusan. 'Investigating the usefulness of i-vectors for automatic language characterization'. In: Speech Prosody 2022 - 11th International Conference on Speech Prosody. Lisbonne, Portugal, 23rd May 2022. DOI: 10.21437/speechprosody.2022-94. URL: https://hal.science/hal-03823002.

[27] P. Tomasello, A. Shrivastava, D. Lazar, P.-C. Hsu, D. Le, A. Sagar, A. Elkahky, J. Copet, W.-N. Hsu, Y. Adi, R. Algayres, T. A. Nguyen, E. Dupoux, L. Zettlemoyer and A. Mohamed. 'STOP: A dataset for spoken task oriented semantic parsing'. In: *IEEE*. SLT-2022 IEEE. Doha, Qatar, 9th Jan. 2023. URL: https://hal.inria.fr/hal-03989829.

Doctoral dissertations and habilitation theses

[28] R. Riad. ‘Automatic speech and language processing for precisionmedicine in Huntington’s disease’. Ecole normale supérieure - ENS PARIS, 1st Apr. 2022. URL: https://hal.inria.fr/tel-03986765.

Reports & preprints

[29] T. Sinha, Z. Bai and J. Cassell. *A Novel Multimodal Approach for Studying the Dynamics of Curiosity in Small Group Learning*. 19th Jan. 2022. DOI: 10.35542/osf.io/rfxwg. URL: https://hal.inria.fr/hal-03536340.

Other scientific publications

[30] M. Lavechin, M. D. Seyssel, L. Gautheron, E. Dupoux and A. Cristia. ‘Reverse engineering language acquisition with child-centered long-form recordings’. In: *Annual Review of Linguistics* 8 (2022), pp. 389–407. URL: https://hal-cnrs.archives-ouvertes.fr/hal-03992325.

11.3 Cited publications

[31] D. A. Ferrucci. ‘Introduction to “this is watson”’. In: *IBM Journal of Research and Development* 56.3.4 (2012), pp. 1–1.

[32] K. He, X. Zhang, S. Ren and 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.

[33] J. Hernández-Orallo, F. Martínez-Plumed, U. Schmid, M. Siebers and D. L. Dowe. ‘Computer models solving intelligence test problems: Progress and implications’. In: *Artificial Intelligence* 230 (2016), pp. 74–107.

[34] B. M. Lake, T. D. Ullman, J. B. Tenenbaum and S. J. Gershman. ‘Building machines that learn and think like people’. In: *arXiv preprint arXiv:1604.00289* (2016).

[35] T. Linzen, E. Dupoux and Y. Goldberg. ‘Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies’. In: *Transactions of the Association for Computational Linguistics* 4 (2016), pp. 521–535.

[36] C. Lu and X. Tang. ‘Surpassing human-level face verification performance on LFW with GaussianFace’. In: *arXiv preprint arXiv:1404.3840* (2014).

[37] S. T. Mueller. ‘A partial implementation of the BICA cognitive decathlon using the Psychology Experiment Building Language (PEBL)’. In: *International Journal of Machine Consciousness* 2.02 (2010), pp. 273–288.

[38] 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 and D. Hassabis. ‘Mastering the game of Go with deep neural networks and tree search’. In: *Nature* 529.7587 (2016), pp. 484–489.

[39] I. Sutskever, O. Vinyals and Q. V. Le. ‘Sequence to sequence learning with neural networks’. In: *Advances in neural information processing systems*. 2014, pp. 3104–3112.

[40] A. M. Turing. ‘Computing machinery and intelligence’. In: *Mind* 59.236 (1950), pp. 433–460.

[41] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu and G. Zweig. ‘Achieving human parity in conversational speech recognition’. In: *arXiv preprint arXiv:1610.05256* (2016).