Orange
Innovation

Research in NLP at Orange

Lina Rojas, Johannes Heinecke, Quentin Brabant, Gwénolé Lecorvé, Géraldine Damnati and Frédéric Herledan

INNOV/DATA-AI/AITT

orange™

Orange : a World-Leading Telecommunications Operator

142,000
employees

8th
largest global telecoms brand

26 countries
in Europe, Africa and the Middle-East

#1
for digital convergence in Europe

Orange Research

1st patents applicant among European operators on its fields of activity

200 new inventions protected by patent on average per year

775 researchers in France and abroad

including **132** PhD and post-doctoral

1 Scientific Council with 8 leading figures from the scientific

Natural Language Processing and Applications

Why?

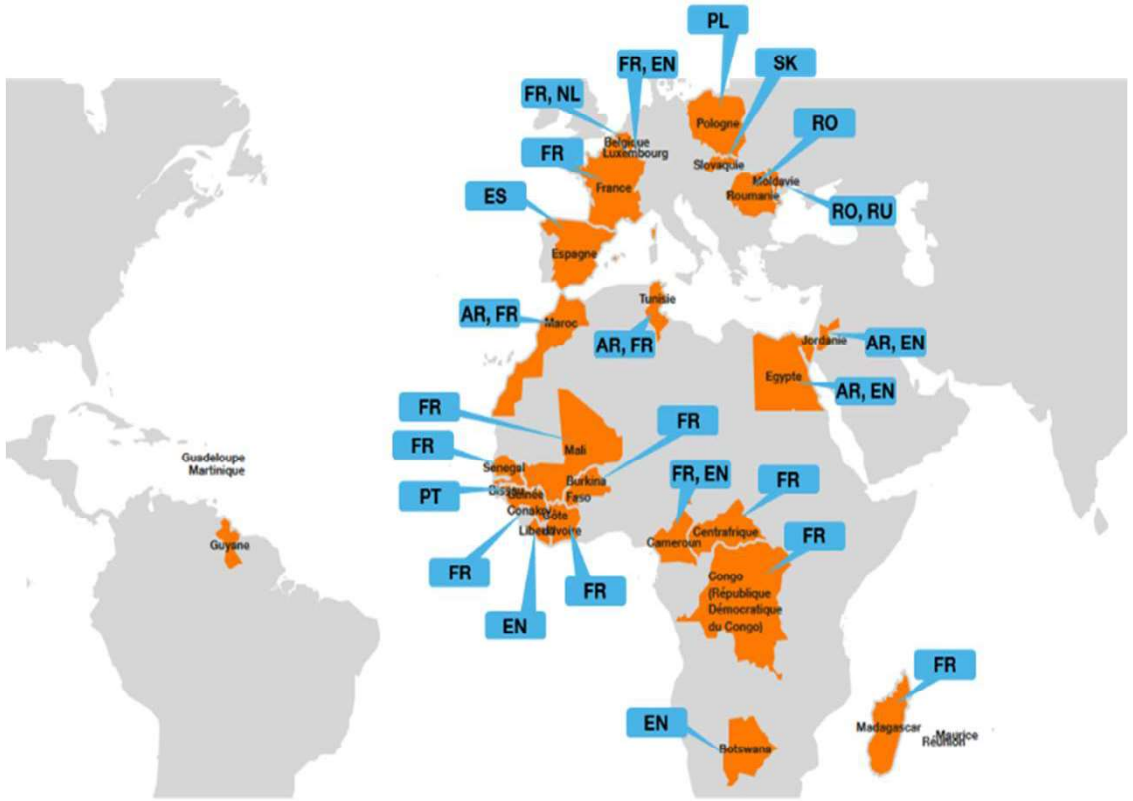
- Big Data
- Data access
- Multilinguality
- Sharing Knowledge

Customer Care

- Call centers
- Chat, SMS, chats
- Customer Reviews
- Technical Assistance
- Net Promotor Score (NPS)

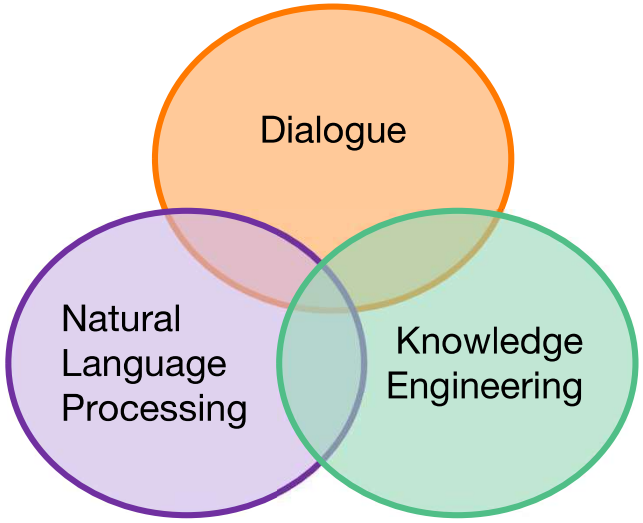
Internal Services

- Document retrieval
- Meetings



Natural Language Processing and Applications

Three scientific axes and numerous tasks linking language and knowledge



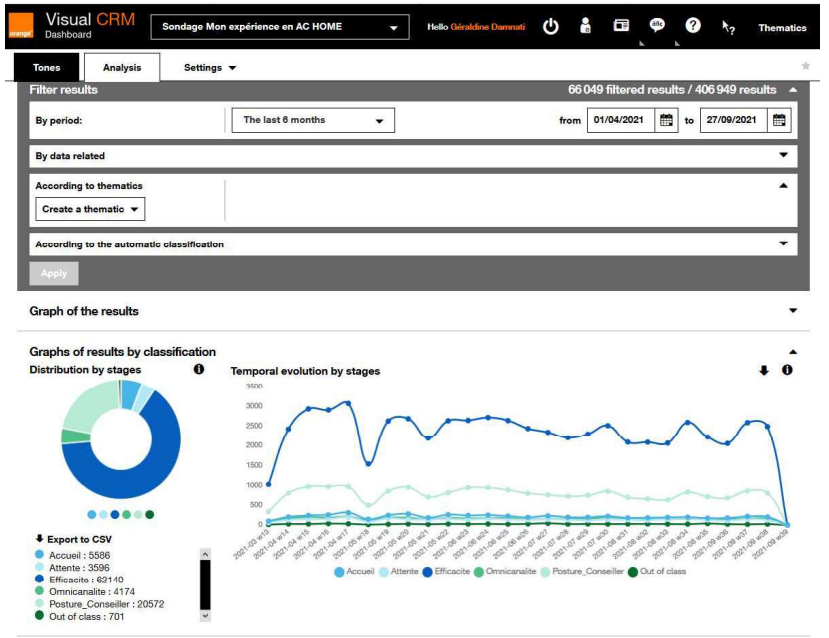
Research subjects

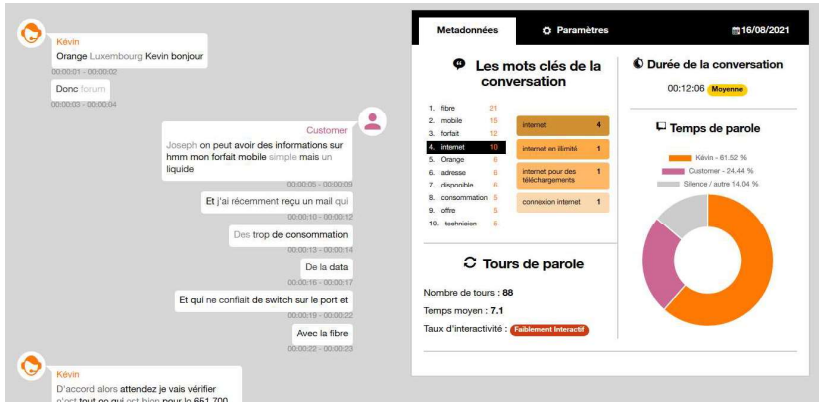
- Multilingual Dependency Syntax parsing
- Multilingual Semantic AMR Parsing
- Information Retrieval
- Reading comprehension
- Information Extraction
- Transform information into (structured) knowledge
- Question Answering using knowledge bases
- Natural Language Understanding and Generation
- Dialogue State Tracking and Dialogue Policy
- Conversational QA

Examples of Applications

- **Text categorization**
  - Multilingual, multitask models, noise resistant (typos, ASR) trained on small datasets to save energy
- **Opinion Analysis**
  - Global tonality + opinion targets + various aspects

  → Technological transfers to operational projects (VisualCRM: used within Orange to classify and explore user comments, feedbacks from surveys towards customers and employees)





Examples of Applications

- **Conversation analytics**
  - Global classification and checkpoints verification
  - Conversation summarization
  - →Transfer to operational projects
- Statistical Dialogue Systems
  - Understanding
  - Dialogue State Tracking
  - Policy Learning
  - Generation
- **Conversational QA and** Reading Comprehension
- End-to-end Approaches

Open Research in NLP

Partnerships

LIS Laboratoire d'Informatique & Systèmes

fondation maison des sciences de l'homme

LABORATOIRE INFORMATIQUE D'AVIGNON
Université d'Avignon et des Pays de Vaucluse

UMR IRISA

EURECOM
Sophia Antipolis

Loria
Laboratoire lorrain de recherche en informatique et ses applications

BnF

European Projects

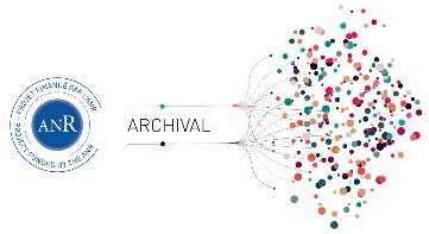
Interactive *Natural Language* Technology for eXplainable Artificial Intelligence
NL4XAI

Conversational Brains

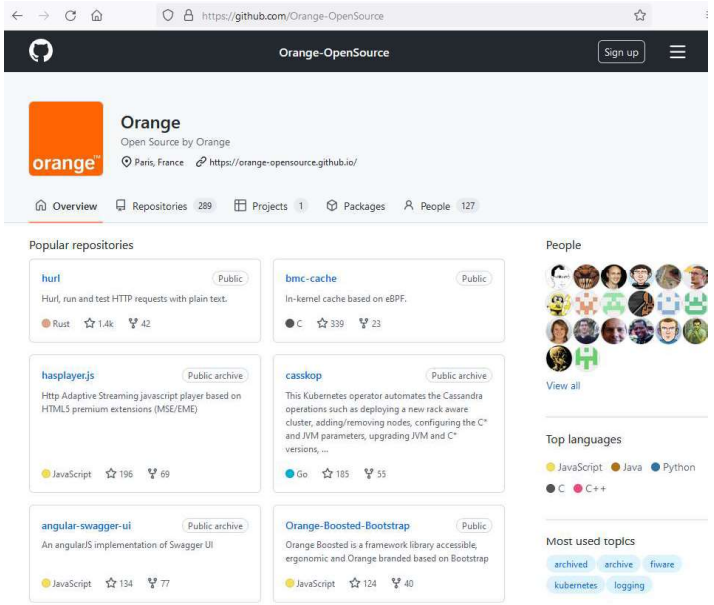
National Project

ANR

ARCHIVAL



Open-source: Code & Corpus



Our papers @LREC 2022

- **Quentin Brabant, Gwenolé Lecorvé, Lina Rojas-Barahona:** ***CoQAR: Question Rewriting on CoQA***
- **Frédéric Béchet, Elie Antoine, Jérémy Auguste, Géraldine Damnati:** ***Question Generation and Answering for exploring Digital Humanities collections***
- **Aline Étienne, Delphine Battistelli, Gwénolé Lecorvé:** ***A (Psycho-)Linguistically Motivated Scheme for Annotating and Exploring Emotions in a Genre-Diverse Corpus***
- **Johannes Heinecke, Anastasia Shimorina:** ***Multilingual AMR for Celtic Languages***

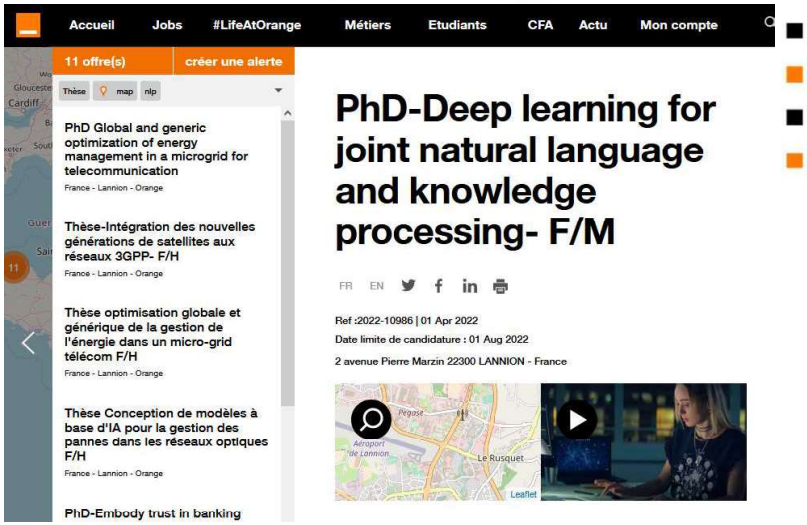
We publish!

- **ACL Conferences and Workshops:**
  - *HEDS-Light: A Lighter Datasheet for Recording Details of Human Evaluation Experiments in NLP*. In: 15th International Conference on Natural Language Generation **INLG 2022**. Maine, USA
  - *Hyperbolic Temporal Knowledge Graph Embeddings with Relational and Time Curvatures.* In: Findings of the Association for Computational Linguistics: **ACL-IJCNLP 2021**",aug,202. Association for Computational Linguistics.
  - *Transfer Learning and Masked Generation for Answer Verbalization*. In: **SUKI@NAACL 2022**
  - *The Human Evaluation Datasheet: A Template for Recording Details of Human Evaluation Experiments in NLP. In Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (**ACL HumEval 2022**), pages 54–75, Dublin, Ireland*
  - *Can we predict how challenging Spoken Language Understanding corpora are across sources, languages and domains?* In: International Workshop on Spoken Dialogue System Technology (**IWSDS 2021**).
  - Quentin Brabant, Lina Rojas-Barahona and Claire Gardent: *Active Learning and Multi-label Classification for Ellipsis and Coreference Detection in Conversational Question-Answering*. In: International Workshop on Spoken Dialogue System Technology (**IWSDS 2021**). Singapore.
  - *SpanAlign: Efficient Sequence Tagging Annotation Projection into Translated Data applied to Cross Lingual Opinion Mining.* In: Proceedings of the Seventh Workshop on Noisy User generated Text (**W NUT 2021**)
  - *Hybrid Enhanced Universal Dependencies Parsing.* **IWPT 2020** Shared Task on Parsing into Enhanced Universal Dependencies
  - *CALOR-QUEST: generating a training corpus for Machine Reading Comprehension models from shallow semantic annotations." MRQA: Machine Reading for Question Answering-Workshop at **EMNLP-IJCNLP***
  - *Spoken conversational search for general knowledge. In **SIGDial 2019**.*
  - DATCHA: Syntactic parsing of chat language in contact center conversation corpus. **SIGDIAL. 2016.**

- **NeurIPS Conference and workshops:**
  - *Budgeted Reinforcement Learning in Continuous State Space.* ***NeurIPS 2019***
  - *Controllable Paraphrase Generation with Multiple Types of Constraints.* ***CtrlGen 2021***
  - *Is the User Enjoying the Conversation? A Case Study on the Impact on the Reward Function.* ***HLDS2020***
  - *Diluted Near-Optimal Expert Demonstrations for Guiding Dialogue Stochastic Policy Optimisation.* ***HLDS2020***
- **LREC :**
  - *Cross-lingual and cross-domain evaluation of Machine Reading Comprehension with Squad and CALOR-Quest corpora.* ***LREC 2020.***
  - *A Multimodal Educational Corpus of Oral Courses: Annotation, Analysis and Case Study* ***LREC2020***
- **Semantic Web, Information Retrieval and Speech Conferences**
  - *DAGOBAH UI: A New Hope For Semantic Table Interpretation*. In: **ESWC 2022**
  - *A Framework for Automatically Interpreting Tabular Data at Orange.* In: Industry Track, 20th International Semantic Web Conference (**ISWC 2021**), remote, USA
  - *Calor-Dial a corpus for Conversational Question Answering on French encyclopedic documents*. In: **CIRCLE 2022.**
  - *BreizhCorpus: a Large Breton Language Speech Corpus and its use for Text-to-Speech Synthesis.* In: **Speech Odyssey 2022**
- **French Conferences and Journals :**
  - *Génération de question à partir d'analyse sémantique pour l'adaptation non supervisée de modèles de compréhension de documents*. In: **TALN 2022**, Avignon
  - *Une chaîne de traitement pour prédire et appréhender la complexité des textes pour enfants d'un point de vue linguistique OU Dans quelle mesure peut-on se fier aux tranches d'âge des éditeurs ?* In: **TALN 2022**, Avignon
  - *Etiquetage ou generation de sequence pour la comprehension automatique du langage en context d'interaction ?* In: **TALN 2022**, Avignon
  - *Génération automatique de texte en langage naturel pour les systèmes de questions-réponses*. In: revue **Traitement Automatique des Langues** 62(1) 2021, pp. 13-37.

Orange
Innovation

Join us →
(orange.jobs)



Thank you!

Data and AI

orange™