Lucía **Ormaechea**

Ph.D. Candidate · NLP Researcher · Computational Linguist

(+41) 78 220 49 79 | lucía-ormaechea-grijalba | luciaormaechea.com | lormaechea
| Lucia.OrmaecheaGrijalba@unige.ch | Lucía Ormaechea

Experience

Research & Teaching Assistant
Université de Genève
Geneva, Switzerland
Dec. 2020 - Present

- Contributed to **BabelDr** ⧉ :
  - Developed and enhanced a **specialized Automatic Speech Recognition (ASR)** system based on Kaldi.
  - Deployed a **multi-container Docker application** to perform ASR inside the **BabelDr medical translation device**.
- Participated in **PROPICTO** ⧉ :
  - Compilation of *complex-simple* **sentence pairs** from comparable corpora to be further used for the **generation of simplified texts**.
  - Development and fine-tuning of cascade and end-to-end models for **automatic simplification of spontaneous French**.
- Collaborated as an assistant in **Localization** ⧉ (M.A. Course): preparation of course materials, assignment grading, assistance and support to students during practical sessions.

NLP Research Intern
Laboratoire d'Informatique de Grenoble (LIG)
Grenoble, France
Feb. 2020 - Sep. 2020

- Participated in BabelDr: development of a **Kaldi-based ASR system for medical-related applications**.
- **Main tasks**:
  - Created a complete pipeline for injecting **grammar-based language models** into the Kaldi Speech Recognition Toolkit.
  - Trained **HMM-DNN-based acoustic models** using open-source corpora.
  - Developed a **Kaldi web-server API** for ASR applications, containerized ASR-related tools using Docker.
  - Conducted a prototype testing and evaluation.

Translation Intern
New York Habitat
Remote Working
Jul. 2018 - Sep. 2018

- Translated **ENG > ESP** commercial texts: video transcriptions – travel articles – apartment reviews – client testimonials.
- **Selected publications**:
  - *Bajo la sombra del cine: una guía del cine en Nueva York (parte 1: Manhattan)* ⧉ , (2018).
  - *El cine de los distritos: un recorrido por el cine de Nueva York (parte 2)* ⧉ , (2018).
  - *Guía de viaje de verano: nuestros mejores planes en Londres* ⧉ , (2018).

Undergraduate Research Assistant
Universidad de Navarra
Pamplona, Spain
Sep. 2016 - Jun. 2017

- Collaborated as a research student at the **Department of Philology** ⧉
- **Main tasks**:
  - Development of educational materials aiming to automatically evaluate knowledge of Spanish language in L2 learners.
  - One-on-one tutorial assessment on academic writing.
  - Proofreading of academic papers, document classification.

Education

Ph.D. (Joint Degree) in Computer Science & Multilingual Information Processing
Université de Genève – Université Grenoble-Alpes
Geneva, Switzerland
Jun. 2022 - present

- Currently pursuing a **joint Ph.D.** between the **Departement of Translation Technology** ⧉ at UNIGE, and the **GETALP team** ⧉ , at UGA.
- **Ph.D. Project**: *Towards Simpler Transcripts: Investigating Automatic Simplification of French Spontaneous Speech* ⧉

M.Sc. in Natural Language Processing
Institut National des Langues et Civilisations Orientales (INALCO) – Université Sorbonne-Nouvelle (Paris 3)
Paris, France
Sep. 2018 - Sep. 2020

- **Graduated with High Honors.**
- **Coursework**: Scripting (Bash, Python, Perl) – Object-Oriented Programming (C++, Java) – Statistical Methods for Corpus Exploitation – Text Mining – Convolutional Neural Networks (Keras) – Corpus Linguistics – Mark-up Languages (XML, XSLT) – Databases (SQL, Neo4J).
- **Master's Thesis**: *Mise en place d'un système robuste de reconnaissance automatique de la parole appliqué au domaine médical* ⧉ **GPA: 19/20**

B.A. in Hispanic Philology
Universidad de Navarra
Pamplona, Spain
Sep. 2014 - Jun. 2018

- **Extraordinary End-of-Degree Award Nominee.**
- **Coursework**: Lexicology – Semantics – Phonetics – Phonology – Morphology – Syntax – Discourse Analysis.

Teaching & Assistantships

2021-2024	**M.A. in Translation, M.A. in Multilingual Communication Technology** – *Teaching Assistant*, Localization, with L. Morado (2021, 2022, 2024), and S. Girletti (2023), Université de Genève
2016-2017	**Workshops on Spanish Language** – *Instructor*, Spanish grammar courses for non-native undergraduate students, Universidad de Navarra

Languages

	Listen	Read	Speak	Write	
Spanish	*Native speaker*				
English	C2	C2	C2	C2	*Advanced Certificate in English (CAE)*
French	C2	C2	C2	C2	*Diplôme Approfondi de Langue Française (DALF)*
Italian	B2	B2	B1	B1	

Software Skills

Programming	Python, Bash (*proficient*) – Perl, SQL (*comfortable*) – C++, Java (*some experience*)
AI Models & Techniques	Supervised learning – LLMs – Pre-training – Fine-tuning (full and parameter-efficient) – Prompt engineering
Libraries	Transformers – Speechbrain – Langchain – Flask – Keras – Fairseq – Pandas – NLTK – SpaCy – Scrapy
Tools	Kaldi – Slurm – Git – Vim – LaTeX – Docker – Praat – Pip – Nginx – Sublime Text – Hugging Face Spaces
Operating Systems	GNU/Linux (Ubuntu, Debian, Linux Mint) – Windows (10, 11) – Android
Web Development	HTML – CSS – Jekyll – XML – XSLT – JavaScript – Gradio – MkDocs – GitHub Pages
Software	Microsoft Office – Gate – Trados – Passolo – Adobe Photoshop – GIMP

Scholarships & Awards

2020	*Candoc* **Research Fellowship** – Granted by the Swiss National Science Foundation (SNSF)
2013	**Runner-up at the 4th Inter-University Debate Tournament** ⮺ – Universidad Pública de Navarra

Selected Publications

2024	**Lucía Ormaechea** and Nikos Tsourakis. *Automatic Text Simplification for French: Model Fine-Tuning for Simplicity Assessment and Simpler Text Generation* ⮺ In: International Journal of Speech Technology ⮺, Springer, vol. 27.
2024	Jonathan Mutal and **Lucía Ormaechea**. *TIM-UNIGE Translation into Low-Resource Languages of Spain for WMT24* ⮺ In: Proceedings of the Ninth Conference on Machine Translation ⮺ (WMT), Miami (USA).
2024	**Lucía Ormaechea**, Nikos Tsourakis, Didier Schwab, Pierrette Bouillon and Benjamin Lecouteux. *Simplification Strategies in French Spontaneous Speech* ⮺ In: Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context ⮺, within the Joint International Conference LREC-COLING ⮺, Torino (Italy).
2023	**Lucía Ormaechea**, Nikos Tsourakis, Didier Schwab, Pierrette Bouillon and Benjamin Lecouteux. *Simple, Simpler and Beyond: A Fine-Tuning BERT-Based Approach to Enhance Sentence Complexity Assessment for Text Simplification* ⮺ In: Proceedings of the 6th International Conference on Natural Language and Speech Processing ⮺ (ICNSLP), Trento (Italy).
2023	**Lucía Ormaechea** and Nikos Tsourakis. *Extracting Sentence Simplification Pairs from French Comparable Corpora Using a Two-Step Filtering Method* ⮺ In: Proceedings of the 8th Swiss Text Analytics Conference ⮺, Neuchâtel (Switzerland).
2023	**Lucía Ormaechea**, Pierrette Bouillon, Maximin Coavoux, Emmanuelle Esperança-Rodier, Johanna Gerlach *et al.* *PROPICTO: Developing Speech-to-Pictograph Translation Systems to Enhance Communication Accessibility* ⮺ In: Proceedings of the 24th Annual Conference of The European Association for Machine Translation ⮺ (EAMT), Tampere (Finland).
2022	**Lucía Ormaechea**, Benjamin Lecouteux, Pierrette Bouillon and Didier Schwab. *A Tool for Easily Integrating Grammars as Language Models into the Kaldi Speech Recognition Toolkit* ⮺ In: ESSLLI 2022 - Workshop on Bridges and Gaps between Formal and Computational Linguistics, Galway (Ireland).

Toolkits & Datasets

2023	**WiViCo (Wikipedia Vikidia Corpus)** ⮺ – A general-purpose parallel sentence simplification dataset for French, automatically extracted by using a two-step filtering method to mine register-diversified comparable corpora.
2021	**kaldi-grammar-compiler** ⮺ – Language modeling tool designed to transform fixed grammars into compiled Finite State Transducers (FSTs). This allows to make them readable as language models by the Kaldi Speech Recognition Toolkit.

Academic Service

2024	*Corpus* Journal No. 26 ⌐ – Managing noise during computer processing of linguistic corpora – **Peer Reviewer**
2023	**Bruit de fond ou valeur ajoutée ?** *Gérer le bruit lors des traitements informatiques des corpus linguistiques* ⌐ – Seminar jointly organized by Université Grenoble-Alpes and Sapienza Università di Roma – **Member of Scientific Committee**
2018	**5th Forum of the Association for Discourse Linguistics** ⌐ – Held at Instituto de Cultura y Sociedad, Universidad de Navarra

Certifications

Getting Started With Mistral ⌐
Online Course

DeepLearning.AI
September 2024

Short course on Mistral's API features to enhance structured LLM responses (JSON mode, function calling and RAG).

3rd Advanced Language Processing School (ALPS) ⌐
Winter School

Université Grenoble-Alpes – Naver Labs Europe
January 2023

Winter school featuring advanced lectures, social and poster sessions covering a wide range of hot topics in NLP.

Research Methods in Translation & Interpreting Studies (ReMeTIS) ⌐
Online Course

Université de Genève
December 2022

• Foundations & Data Analysis: qualitative and mixed methods, research ethics.
• Specific Research & Scientific Communication Skills: corpus construction, quantitative data analysis, publication strategies.

Sequence Models ⌐
Online Course

Coursera, DeepLearning.AI
November 2022

Deep Learning Specialization (V): Recurrent Neural Networks – Long Short Term Memory – Attention Models.

Structuring Machine Learning Projects ⌐
Online Course

Coursera, DeepLearning.AI
October 2022

Deep Learning Specialization (III): Error Analysis in Machine Learning Systems – Performance Improvement Strategies.

Improving Deep Neural Networks: Hyperparameter Tuning, Regularization & Optimization ⌐
Online Course

Coursera, DeepLearning.AI
October 2022

Deep Learning Specialization (II): Optimization Algorithms – Hyperparameter Tuning.

Neural Networks & Deep Learning ⌐
Online Course

Coursera, DeepLearning.AI
October 2022

Deep Learning Specialization (I): Deep Learning Foundations – Neural Network Architectures.

33rd European Summer School in Logic, Language & Information (ESSLLI) ⌐
Summer School

National University of Ireland Galway
August 2022

Interdisciplinary summer school involving courses and workshops in logic, linguistics and computer science.

2nd Advanced Language Processing School (ALPS) ⌐
Winter School

Université Grenoble-Alpes – Naver Labs Europe
January 2022

Winter school featuring advanced lectures, social and poster sessions covering a wide range of hot topics in NLP.

Personal Data Protection ⌐
MOOC Course

Institut National de Recherche en Sciences et Technologies du Numérique (Inria)
May 2020

MOOC Course introducing data confidentiality regulations and privacy protection.

Introduction to Corpus Linguistics ⌐
MOOC Course

Université Grenoble-Alpes (UGA)
June 2019

MOOC Course introducing corpus linguistics and its many implementations in lexicology, discourse analysis and NLP.

Other Achievements

Projet Boîte à Outils ⌐
Master's Project

Institut National des Langues et Civilisations Orientales (INALCO) / Université Sorbonne-Nouvelle (Paris 3)
May 2019

A quantitative research focused on morphosyntactic patterns found on the French newspaper Le Monde.

Projet Documents Structurés ⌐
Master's Project

Institut National des Langues et Civilisations Orientales (INALCO) / Université Sorbonne-Nouvelle (Paris 3)
May 2019

A set of programming exercises related to XML and information extraction.