

alpöktem

contact

Barcelona, Spain

+34 638 40 80 98

alp@oktem.me

alp.oktem@clearglobal.org

<http://alp.oktem.me>

[LinkedIn://alp-oktem](https://www.linkedin.com/company/alp-oktem)

[Github://alpoktem](https://github.com/alpoktem)

[ORCID://0000-0002-0700-1159](https://orcid.org/0000-0002-0700-1159)

languages

turkish native

english fluent

spanish fluent

catalan beginner

programming

Python, Linux,

Processing, MATLAB,

Raspberry Pi

libraries

PyTorch, Docker,

OpenNMT, FastAPI

interests

speech technology,

low-resource NLP,

language preservation,

community-centered AI,

open source, social impact

hobbies

biking, hiking,

snowboard, guitar, drums

experience

2025 - Now

LANGUAGE TECHNOLOGY CONSULTANT

Building accessible AI systems that preserve linguistic diversity and empower under-represented communities

- RAPNIC: Technical lead developing first Catalan impaired speech corpus
- InteractIA: Voice-based chatbots that enable interactive learning and normalize social use for children
- Awal: Creation of open-source models and data for Tamazight MT and ASR via community data collection platform awaldigital.org

2017 - Now

CLEAR GLOBAL/TRANSLATORS WITHOUT BORDERS

Remote

Computational Linguist

Lead organization's expansion into language technology as technical expert, guiding data collection, AI development and product prototyping for crisis and development

- Published 20+ datasets and models, including ASR, TTS and translation systems for 15+ underrepresented African and Asian languages with community support
- Built and integrated production-level MT and Speech APIs for language services
- Co-authored influential publications with [TICO-19](#), [Masakhane](#), and [BibleTTS](#)

2017 - 2025

COL-LECTIVAT SCCL

Barcelona, SPAIN

Co-founder / Technology lead

Pioneering NLP solutions for social impact, focusing on Catalonia and Spain's minoritized, endangered and historical languages through community-powered initiatives

- Spearheaded digitalization initiatives for [Aranese](#), [Tamazight](#) and [Judeo-Spanish](#)
- Pioneered open source neural TTS for [Catalan](#), [Galician](#) and [Judeo-Spanish](#)
- Openly published 12 datasets and 10 models for 5 low-resource languages
- Launched MT apps for [Judeo-Spanish](#), [Tamazight](#) (160k+ requests since 2022)
- Designed and delivered 8 AI and MT workshops to 100+ participants

2014 - 2019

UNIVERSITAT POMPEU FABRA

Barcelona, SPAIN

Researcher in Natural Language Processing Group (TALN)

EU project KRISTINA: A multilingual and interactive digital healthcare assistant

- Developed punctuation restoration for transcribed speech
- Created and annotated the first Turkish surface and deep syntax corpus

Associate Professor

Data structures and algorithms II, Software engineering for web applications

2012 - 2014

FRAUNHOFER IAIS

Sankt Augustin, GERMANY

Student Research Assistant for EU project LinkedTV

education

2014 - 2019

PhD, Computational Linguistics

Universitat Pompeu Fabra, Barcelona, SPAIN

Thesis: Incorporating Prosody into Neural Speech Processing Pipelines - Applications on automatic speech transcription and spoken language machine translation

- Pioneered machine dubbing using prosodic phrase alignment, cited 35+ times
- Created *Heroes Corpus*: first dubbed movie speech corpus for SLT research
- Published software suite for annotating and visualizing prosody in speech data
- Developed novel punctuation restoration for transcribed speech using prosody
- Received 2018 *María de Maeztu DTIC-UPF Open Science Award*

2011 - 2014

MSc, Computer Science

University of Bonn, Bonn, GERMANY

Thesis: Person identification and topic segmentation for news broadcast using automatic speaker clustering and character recognition

2006 - 2010

BsC, Computer Science

Bilkent University, Ankara, TURKEY

Senior year project: Line and Word Segmentation in Historical Manuscripts

2008 - 2009

ERASMUS Exchange student

University of Bath, Bath, UNITED KINGDOM

publications

*full list can be found at:
alp.oktem.me/publications/

Correcting the Tamazight Portions of FLORES+ and OLDI Seed Datasets

Alp Öktem, Mohamed Aymane Farhi, Brahim Essaidi, Naceur Jabouja and Farida Boudichat
November, 2025 - Tenth Conference on Machine Translation (WMT 2025), Suzhou, China

Awal – Community-Powered Language Technology for Tamazight

Alp Öktem, Farida Boudichat
Dec., 2025 - Technologies d'Information et de Communication pour l'AMazighe, Rabat, Morocco

Nós-TTS : a Web User Interface for Galician Text-to-Speech

Carmen Magariños, Alp Öktem and 6 other researchers from Instituto da Lingua Galega
March, 2024 - PROPOR, Santiago de Compostela, Spain

BibleTTS - a large, high-fidelity, multilingual, and uniquely African speech corpus

20 authors from Coqui.ai, Masakhane and various academic institutions
September, 2022 - Interspeech, Seoul, South Korea

Preparing an endangered language for the digital age - The Case of Judeo-Spanish

Alp Öktem, Rodolfo Zevallos, Yasmin Moslem, Özgür Güneş Öztürk, Karen Gerson Şarhon
June, 2022 - EURALI workshop organized within LREC, Marseille, France

Corpora compilation for prosody-informed speech processing

Alp Öktem, Mireia Farrús, Antonio Bonafonte
September, 2021 - Language Resources and Evaluation (Journal paper)

Congolese Swahili Machine Translation for Humanitarian Response

Alp Öktem, Eric DeLuca, Rodrigue Bashizi, Eric Paquin, Grace Tang
April, 2021 - Africa NLP workshop organized within EACL, Online

Received best paper award

TICO-19 – The Translation Initiative for COvid-19

18 authors from George Mason, Carnegie Mellon, John Hopkins University, Translated, Amazon AI, Microsoft, Facebook AI, Appen, Google and Translators without Borders
November, 2020 - NLP COVID-19 Workshop organized within EMNLP, Online

Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages

48 authors from Masakhane
November, 2020 - Findings of EMNLP, Online

Received 2021 Wikimedia Foundation Research of the Year Award

Gamayun – Language Technology for Humanitarian Response

Alp Öktem, Muhannad Albayk Jaam, Eric DeLuca, Grace Tang
October, 2020 - IEEE Global Humanitarian Technology Conference (GHTC), Online

CATOTRON – A Neural Text-to-Speech System in Catalan

Baybars Külebi, Alp Öktem, Alex Peiró-Lilja, Santiago Pascual, Mireia Farrús
October, 2020 - Interspeech, Shanghai, China (Online)

Masakhane – Machine Translation For Africa

25 authors from Masakhane
April, 2020 - AfricaNLP Workshop organized within ICLR, Addis Ababa, Ethiopia (Online)

Tigrinya Neural Machine Translation with Transfer Learning for Humanitarian Response

Alp Öktem, Mirko Plitt, Grace Tang
April, 2020 - AfricaNLP Workshop organized within ICLR, Addis Ababa, Ethiopia (Online)

Prosodic phrase alignment for machine dubbing

Alp Öktem, Mireia Farrús, Antonio Bonafonte
September, 2019 - Interspeech, Graz, Austria

Building an open source automatic speech recognition system for Catalan

Baybars Külebi, Alp Öktem
November, 2018 - IberSPEECH, Barcelona, Spain

Bilingual prosodic dataset compilation for spoken language translation

Alp Öktem, Mireia Farrús, Antonio Bonafonte
November, 2018 - IberSPEECH, Barcelona, Spain

Attentional parallel RNNs for generating punctuation in transcribed speech

Alp Öktem, Mireia Farrús, Leo Wanner
October, 2017 - Statistical Language and Speech Processing (SLSP), Le Mans, France

Prosograph: A tool for prosody visualisation of large speech corpora

Alp Öktem, Mireia Farrús, Leo Wanner
August, 2017 - Interspeech, Stockholm, Sweden