

# alpöktem

## contact

alp@collectivat.cat  
alp.oktem@clearglobal.org

<http://alpoktem.github.io>

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

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

[Twitter://OktemAlp](https://twitter.com/OktemAlp)

## languages

turkish native

english fluent

spanish fluent

## programming

Python, Linux bash,  
Processing, MATLAB

## interests

speech technology,  
computational linguistics,  
machine translation,  
prosody, deep neural  
networks

## hobbies

illustration, thai boxing,  
playing drums and guitar,  
biking, hiking

## experience

2019 - Now **CLEAR GLOBAL/TRANSLATORS WITHOUT BORDERS** Remote

*Computational Linguist*

- Low-resource language dataset preparation: parallel text, voice
- Machine translation engine development for Arabic, Hausa, Swahili, Tigrinya
- Speech recognition engine development for Bengali, Congolese Swahili
- Language tech consultancy for chatbot, data collection, telephone IVR projects

2017 - Now **COL-LECTIVAT SCCL** Barcelona, SPAIN

*Co-founder / Computational Linguistics Expert*

- Collection of open speech (ASR and TTS) datasets for Catalan and Ladino
- Machine translation engine development for Ladino
- Organising applied language technology workshops

2014 - 2019 **UNIVERSITAT POMPEU FABRA** Barcelona, SPAIN

*Associate Professor*

Data structures and algorithms II, Software engineering for web applications

2014 - 2018 *Researcher in Natural Language Processing Group (TALN)*

Work carried out for EU project *KRISTINA: A multilingual and interactive digital healthcare assistant*

- Developing a module for punctuation restoration of raw transcribed speech input
- Preparing and annotating a Turkish surface and deep syntax corpus

2012 - 2014 **FRAUNHOFER IAIS** Bonn, GERMANY

*Student Research Assistant*

Work carried out for EU project *LinkedTV: Television linked to the web*

- Automatic generation of a speaker identification dataset from raw TV excerpts
- Automatic topic segmentation in news shows using word semantic distances

## education

Universitat Pompeu Fabra, Barcelona, SPAIN

2014 - 2019 **PhD, Computational Linguistics**

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

- Open-source software suite for collecting, handling, prosodic annotation and visualization of monolingual and bilingual speech data
- An English-Spanish parallel movie speech corpus
- Spoken language punctuation restoration with prosodic cues
- Prosody-informed neural machine translation

Awards:

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 banner text recognition*

Other course projects:

- *Score informed audio source separation*, decomposing right and left hand notes in piano recordings using score information from MIDI
- German to English statistical machine translation

2006 - 2010 **BsC, Computer Science** Bilkent University, Ankara, TURKEY

Senior project: *Line and Word Segmentation in Historical Manuscripts*

Skew tolerant segmentation of text lines and words from scanned images of M.K. Atatürk's handwritings and Ottoman manuscripts

2008 - 2009 **ERASMUS Exchange student** University of Bath, Bath, UNITED KINGDOM

## publications

### *Corpora compilation for prosody-informed speech processing*

Alp Öktem, Mireia Farrús, Antonio Bonafonte  
September, 2021 - Language Resources and Evaluation

### *Congolese Swahili Machine Translation for Humanitarian Response*

Alp Öktem, Eric DeLuca, Rodrigue Bashizi, Eric Paquin, Grace Tang  
April, 2021 - Africa NLP workshop organized within the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL2021), Online

### *TICO-19 – The Translation Initiative for COvid-19*

18 authors from various organisations  
November, 2020 - NLP COVID-19 Workshop (Part 2) @ EMNLP 2020, Online

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

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

### *Gamayun – Language Technology for Humanitarian Response*

Alp Öktem, Muhannad Albayk Jaam, Eric DeLuca, Grace Tang  
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 Eighth International Conference on Learning Representations (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 Eighth International Conference on Learning Representations (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

### *Visualizing punctuation restoration in speech transcripts with Prosograph*

Alp Öktem, Mireia Farrús, Antonio Bonafonte  
September, 2018 - Interspeech, Hyderabad, India

### *Attentional parallel RNNs for generating punctuation in transcribed speech*

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

### *Revising the METU-Sabancı Turkish Treebank: An exercise in surface-syntactic annotation of agglutinative languages*

Alicia Burga, Alp Öktem, Leo Wanner  
September, 2017 - International Conference on Dependency Linguistics (Depling), Pisa, Italy

### *Prosograph: A tool for prosody visualisation of large speech corpora*

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

### *Automatic extraction of parallel speech corpora from dubbed movies*

Alp Öktem, Mireia Farrús, Leo Wanner  
August, 2017 - 10th Workshop on Building and Using Comparable Corpora (BUCC), Vancouver, Canada

### *From raw data to semantically enriched hyperlinking: Recent advances in the LinkedTV analysis workflow*

D. Stein, A. Öktem, E. Apostolidis, V. Mezaris, J.-L. Redondo García, R. Troncy, M. Sahuguet, B. Huet  
October, 2013 - Proceedings of Networked & Electronic Media (NEM) Summit, Nantes, France

### *Processing the manuscripts of Atatürk*

T. Soysal, H. Adiguzel, A. Öktem, A. Haman, E.F. Can, P. Duygulu, M. Kalpakli  
April, 2010 - IEEE 18th Signal Processing and Communications Applications Conference (SIU), Diyarbakir, Turkey