Marco Cognetta

cognetta.marco@gmail.com
mcognetta.github.io

Education

2022-Present **PhD Computer Science**, *Tokyo Institute of Technology*, Tokyo, Japan
- Advisor: Dr. Naoaki Okazaki
- MEXT Scholar (文部科学省奨学金)
- Expected Graduation Fall 2025

2016-2018 **MS Computer Science**, *Yonsei University*, Seoul, South Korea
- Advisor: Dr. Yo-Sub Han
- Thesis: *Efficient Algorithms for Two Parsing Problems on Probabilistic Finite Automata*

2011-2015 **BS Discrete Mathematics**, *Georgia Institute of Technology*, Atlanta, GA
- Minor in Korean
- Advisor: Dr. Anton Leykin
- Thesis: *Straight Line Programs and Automatic Differentiation in Python*

Work Experience

05/22-Present **PhD Student Researcher**, *Google*, Tokyo, Japan
- Researcher focusing on language modeling and federated analytics on the Gboard team.
- Developing efficient decoders for on-device language models.

08/19-04/22 **Software Engineer**, *Google*, Mountain View, CA
- Developed compressed statistical language models for mobile keyboard input.
- Developed TensorFlow Federated infrastructure and compressed sketching models for privacy-preserving federated analytics.
- TA for Google Tech Exchange Applied Data Structures and Algorithms (2020, 2021).

02/19-05/19 **Software Engineering Intern**, *Google*, New York City, NY
- Developed a compression algorithm for finite-state transducers used in keyboard language models to reduced space requirements by >90% over the uncompressed version and >58% over the prior production compression scheme (published as a first-author paper).

09/16-12/18 **Graduate Teaching Assistant**, *Yonsei University*, Seoul, South Korea

01/16-05/16 **Upper School Computer Science Teacher**, *Maclay School*, Tallahassee, FL

05/15-08/15 **Data Science Intern**, *AirSage Inc.*, Atlanta, GA

05/12-08/12 **Software Development Intern**, *AirSage Inc.*, Atlanta, GA

Skills

- **Programming Languages:** Python, C++, Julia
- **Tools:** PyTorch, Flux.jl, OpenFst, TensorFlow Federated
- **Human Languages:** English, Korean, Esperanto

Publications (*denotes primary authorship)

1. **Tokenization as Finite-State Transduction**
   - *Marco Cognetta**, Naoaki Okazaki. Computational Linguistics.
2. **The `bread emoji` Team's Submission to the 2025 FedCSIS Predicting Chess Puzzle Difficulty Challenge**
   - Tyler Woodruff, Luke Imbing, *Marco Cognetta*. FedCSIS 2025.
   - Our submission placed 2$^{\text{nd}}$ (out of 73 teams) and won a $500 USD prize.
3. **Pitfalls, Subtleties, and Techniques in Automata-Based Subword-Level Constrained Generation**
   - *Marco Cognetta**, David Pohl*, Junyoung Lee, Naoaki Okazaki. TokShop 2025.
4. **Jamo-Level Subword Tokenization in Low-Resource Korean Machine Translation**
   - Junyoung Lee*, *Marco Cognetta**, Sangwhan Moon, Naoaki Okazaki. LoResMT 2025.
5. **Distributional Properties of Subword Regularization**
   - *Marco Cognetta**, Vilém Zouhar, Naoaki Okazaki. EMNLP 2024.
6. **The `bread emoji` Team's Submission to the IEEE BigData 2024 Cup: Predicting Chess Puzzle Difficulty Challenge**
   - Tyler Woodruff, Oleg Filatov, *Marco Cognetta*. IEEE Big Data 2024.
   - Our submission won 1$^{\text{st}}$ place (out of 143 teams) and a $1000 USD prize.
7. **Two Counterexamples to *Tokenization and the Noiseless Channel***
   - *Marco Cognetta**, Vilém Zouhar, Sangwhan Moon, Naoaki Okazaki. LREC-COLING 2024.
8. **Parameter-Efficient Korean Character-Level Language Modeling**
   - *Marco Cognetta**, Sangwhan Moon, Lawrence Wolf-Sonkin, Naoaki Okazaki. EACL 2023.

9. **SoftRegex: Generating Regex from Natural Language Descriptions using Softened Regex Equivalence**
   - Jun-U Park, Sang-Ki Ko, *Marco Cognetta*, Yo-Sub Han. EMNLP 2019.
10. **On the Compression of Lexicon Transducers**
   - *Marco Cognetta**, Cyril Allauzen, Michael Riley. FSMNLP 2019.
11. **Online Infix Probability Computation for Probabilistic Finite Automata**
   - *Marco Cognetta**, Yo-Sub Han, Soon Chan Kwon. ACL 2019.
12. **Incremental Computation of Infix Probabilities for Probabilistic Finite Automata**
   - *Marco Cognetta**, Yo-Sub Han, Soon Chan Kwon. EMNLP 2018.
13. **Online Stochastic Pattern Matching**
   - *Marco Cognetta**, Yo-Sub Han. CIAA 2018.

Preprints (*denotes primary authorship)

1. **Tutorial: φ-Transductions in OpenFst via the Gallic Semiring**
   - *Marco Cognetta**, Cyril Allauzen.
   - `https://arxiv.org/abs/2506.17942`
2. **An Analysis of BPE Vocabulary Trimming in Neural Machine Translation**
   - *Marco Cognetta**, Tatsuya Hiraoka, Naoaki Okazaki, Rico Sennrich, Yuval Pinter.
   - `https://arxiv.org/abs/2404.00397`

Conference Talks

1. **LotteryTickets.jl: Sparsify Your Flux Models**
   - JuliaCon 2023 (Boston, USA)
   - `https://www.youtube.com/watch?v=ZmcaUyZLi4Q`
   - `https://github.com/mcognetta/LotteryTickets.jl`

Invited Talks

1. **Subword Tokenization Meets Formal Language Theory**
   - Invited Tutorial at Developments in Language Theory (DLT) (August 2025, Seoul)
   - `https://github.com/mcognetta/subword_tokenization_meets_formal_language_theory`
2. **The Tokenization Landscape**
   - National Institute of Advanced Industrial Science and Technology Artificial Intelligence Research Center's Knowledge and Information Research Team (AIST AIRC-KIRT) (March 2024, Tokyo)

Service

- **Women in Science Japan's Machine Learning Summer School for Scientists** - Mentor (2025)
- **Seminars on Formal Languages and Neural Networks (FLaNN)** - Co-organizer (2022 - 2025)
- **Workshop for Natural Language Processing Open Source Software (NLP-OSS)** - Programme Committee (2023)
- **The Gradient** (`https://thegradient.pub/`) - Editorial Board (2021 - 2024)
- **FSU ACM Programming Contest** - Question Writer (2020, 2021$^{(\times 2)}$, 2022)
- **Hackbright Academy** - Volunteer Mentor (2020$^{(\times 2)}$, 2021)
- **ACM International Collegiate Programming Contest (Korea Regional)** - Question Writer (2017, 2018)

Advising

1. **Gordon Lichtstein** (2024) - *Esperanto Morphological Tokenization*
   - High School Extracurricular Senior Project
2. **Junyoung Lee** (2023) - *Jamo-Level BPE in Korean Machine Translation*
   - Nanyang Technological University (NTU Singapore) Bachelor's Thesis
   - Accepted as a full paper at LoResMT 2025
3. **Emil Hukic** (2022) - *FST Tokenization for NLP*
   - Young Science and Engineering Researchers Program (YSEP) Final Project
4. **Kosuke Endo** (2022) - 画像キャプション生成におけるJPEG圧縮への頑健性の改善
   - English Title: *Improved Robustness to JPEG Compression in Image Caption Generation*
   - Tokyo Institute of Technology Bachelor's Thesis (co-advised with Zhishen Yang)
   - Presented at the Japanese Association for Natural Language Processing conference (NLP2023)
5. **Haksu Kim, Yumin Lim, Myeongjang Pyeon** (2018) - *Solving k-MPS using Probabilistic Finite-State Automata*
   - Yonsei University Capstone Project