

Chihiro Taguchi

Updated June 15, 2024

田口 智大

✉ **Email:** ctaguchi@nd.edu

🐙 **GitHub:** [ctaguchi](https://github.com/ctaguchi)

🌐 **Website:** <https://ctaguchi.github.io>

Education

**College of Engineering,
The University of Notre Dame** IN, United States
PhD in Computer Science and Engineering. Aug 2022 – Jul 2027

**School of Philosophy, Psychology and Language Sciences,
The University of Edinburgh** Edinburgh, United Kingdom
MSc by Research in Linguistics with Distinction. Sep 2021 – Aug 2022

**Division of Information Science,
Nara Institute of Science and Technology** Nara, Japan
MA in Engineering. Oct 2020 – Sep 2022

**School of Oriental and African Studies,
The University of London** London, United Kingdom
Exchange Student in Linguistics. Sep 2018 – Jun 2019

**Department of Political Science, Faculty of Law,
Keio University** Tokyo, Japan
Bachelor of Laws (LL.B.). Apr 2015 – Sep 2019

Training

Linguistic Society of America (LSA) Institute. Jun – Jul 2023
University of Massachusetts, Amherst, MA, United States.

Summer Language Abroad Program (Quechua), University of Notre Dame. May – Jun 2023
Quito, Ecuador.

Summer School in Languages and Linguistics. Jul 2022
Leiden University, Netherlands.

Advanced Core Training in Linguistics (ACTL) Summer School. Jun 2022
University of York, United Kingdom.

Lithuanian Language Summer Course. Jul 2019
Vilnius University, Lithuania.

Summer School in Languages and Linguistics. Jul 2018
Leiden University, Netherlands.

3rd Miyako Language Documentation Training Workshop. Feb 2018
Ikemajima Island, Okinawa, Japan.

	Summer School in Languages and Linguistics. Leiden University, Netherlands.	Jul 2017
	Armenian Language Summer School. ARMACAD, Armenia.	Jul – Aug 2016
Experience	Program Committee. 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI) at LREC–COLING 2024.	Dec 2023 – Present
	Student Researcher. AnalytiXIN.	Aug 2023 – Dec 2023
	Research Intern. Pluszero Co.	Apr 2021
	Research Assistant. National Institute of Informatics, Japan.	Apr 2021 – Sep 2021
	Cookpad Online Spring Internship 2021. Cookpad Inc.	Mar 2021
Publications	Language Complexity and Speech Recognition Accuracy: Orthographic Complexity Hurts, Phonological Complexity Doesn't. Chihiro Taguchi , David Chiang. <i>ACL 2024</i> , to appear. Preprint: https://arxiv.org/abs/2406.09202	
	Japanese Rule-based Grapheme-to-phoneme Conversion System and Multilingual Named Entity Dataset with International Phonetic Alphabet. Yuhi Matogawa, Yusuke Sakai, Taro Watanabe, Chihiro Taguchi . <i>SIGMORPHON 2024</i> at NAACL, to appear.	
	KILLKAN: The Automatic Speech Recognition Dataset for Kichwa with Morphosyntactic Information Chihiro Taguchi , Jefferson Saransig, Dayana Velásquez, David Chiang. <i>LREC-COLING 2024</i> , 2024. Paper: https://aclanthology.org/2024.lrec-main.852/	
	J-SNACS: Adposition and Case Supersenses for Japanese Joshi. Tatsuya Aoyama, Chihiro Taguchi , Nathan Schneider. <i>LREC-COLING 2024</i> , 2024. Paper: https://aclanthology.org/2024.lrec-main.839	
	Strategies for the Annotation of Pronominalised Locatives in Turkic Universal Dependency Treebanks. Jonathan Washington, Çağrı Çöltekin, Furkan Akkurt, Bermet Chontaeva, Soudabeh Eslami, Gulnura Jumaliev, Aida Kasieva, Aslı Kuzgun, Büşra Marşan, Chihiro Taguchi . <i>Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024)</i> at LREC-COLING 2024, 2024. Paper: https://aclanthology.org/2024.lrec-main.839	
	Non-discourse-configurationality in Imbabura Kichwa. Chihiro Taguchi , Jefferson Saransig. <i>Proceedings of the Linguistic Society of America</i> , 2024. Paper: https://journals.linguisticsociety.org/proceedings/index.php/PLSA/article/view/5687	

正書法および音韻の複雑さによる音声認識の精度への影響 [The effect of orthographic and phonological complexities on automatic speech recognition accuracy].

Chihiro Taguchi. 第 30 回言語処理学会年次大会 [The 30th Annual Meeting of the Association of Natural Language Processing (NLP2024)], 2024. Paper: https://www.anlp.jp/proceedings/annual_meeting/2024/pdf_dir/B5-1.pdf

Unifying the Annotations in Turkic Universal Dependencies Treebanks.

Çağrı Çöltekin, Furkan Akkurt, Olcay Taner Yildiz, Sardana Ivanova, Tunga Gungor, Suzan Uskudarli, Mehmet Oguz Derin, Aida Kasieva, Gulnura Dzhumalieva, Bermet Chontaeva, Jonathan Washington, Soudabeh Eslami, **Chihiro Taguchi**, Asli Kuzgun, Büşra Marşan, Murat Jumashev, Balkız Öztürk. UniDive 2024 General Meeting, 2024.

Universal Automatic Phonetic Transcription into the International Phonetic Alphabet

Chihiro Taguchi, Yusuke Sakai, Parisa Haghani, David Chiang. *INTERSPEECH 2023*, 2023. ArXiv: <https://arxiv.org/abs/2308.03917>

形態論情報付き日本語 **Universal Dependencies** [Japanese Universal Dependencies with morphological information].

Chihiro Taguchi, So Miyagawa. 第 29 回言語処理学会年次大会 [The 29th Annual Meeting of the Association of Natural Language Processing (NLP2023)], 2023. Paper: https://www.anlp.jp/proceedings/annual_meeting/2023/pdf_dir/P3-6.pdf

沖縄語の **Universal Dependencies** ツリーバンクコーパスの構築 [Building a treebank corpus for Okinawan Universal Dependencies].

So Miyagawa, Hiroshi Kanayama, **Chihiro Taguchi**, Nana Tohyama. 第 29 回言語処理学会年次大会 [The 29th Annual Meeting of the Association of Natural Language Processing (NLP2023)], 2023. https://www.anlp.jp/proceedings/annual_meeting/2023/pdf_dir/P3-8.pdf

Introducing Morphology in Universal Dependencies Japanese.

Chihiro Taguchi, David Chiang. *The Sixth Workshop on Universal Dependencies (UDW)*, 2023. Paper: <https://aclanthology.org/2023.udw-1.8>

Mermaid Construction in Lexical Functional Grammar.

Chihiro Taguchi. *Proceedings of the LFG'22 Conference*, 2022. Paper: <https://lfg-proceedings.org/lfg/index.php/main/article/view/19>

名詞述語を伴う繰り上げ構文としての人魚構文 [Mermaid Construction as raising with a nominal predicate] (in Japanese).

Chihiro Taguchi. *Proceedings of the 164th Conference of the Linguistic Society of Japan*, 2022. Paper: https://ls-japan.jpn.org/modules/documents/LSJpapers/meeting/164/handouts/b/B-1_164.pdf

Universal Dependencies Treebank for Tatar: Incorporating Intra-Word Code-Switching Information.

Chihiro Taguchi, Sei Iwata, Taro Watanabe. *Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI)* at LREC 2022, 2022. Paper: <https://aclanthology.org/2022.eurali-1.17>

Mermaid Construction in Tatar.

Chihiro Taguchi. *Proceedings of the 162nd Conference of the Linguistic Society of Japan*, 2021. Paper: https://ls-japan.jp/modules/documents/LSJpapers/meeting/162/handouts/c/C-3_162.pdf

Transliteration for Low-Resource Code-Switching Texts: Building an Automatic Cyrillic-to-Latin Converter for Tatar.

Chihiro Taguchi, Yusuke Sakai, Taro Watanabe. *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching* at NAACL 2021, 2021. Link: <https://aclanthology.org/2021.calcs-1.18>

Chihiro Taguchi, Taro Watanabe. So-Called “Prepositions” in Somali are Not Prepositions: A Linguistic Approach for Somali POS Tagging. *Proceedings of the Annual Meeting of the Association of Natural Language Processing*, 2021.

Chihiro Taguchi. Raising to Quirky Subject in Tatar. *Proceedings of the 161st Conference of the Linguistic Society of Japan*, 2020.

Chihiro Taguchi. Vowel Harmonies in Kazan Tatar: Element Theoretical Approach. *Proceedings of the 159th Conference of the Linguistic Society of Japan*, 2019.

Chihiro Taguchi. 近現代における言語衰退過程の観察と考察 – 琉球語宮古池間方言を事例に – [Observations of the Process of Language Extinction in the Modern Society: A Case Study of the Ikema Dialect of Miyako Ryukyuan] (in Japanese). *政治学研究 [Studies in Political Science]*, 2018. Keio University Press.

Projects

科研費基盤研究 (C) ジンポー語口承資料の資源化に関する研究.

(Grants-in-Aid for Scientific Research: “Digitization of Jinghpaw oral literature”) 2024 – Present
In collaboration with Dr. Keita Kurabe.
日本学術振興会 [Japan Society for the Promotion of Science (JSPS)].

ユーラシア諸言語における談話の研究.

(Studies on the grammar of discourse in Eurasian languages) Nov 2023 – Present
In collaboration with Dr. Norikazu Kogura (PI), Dr. Jakshylyk Akmatallieva, Dr. Fuyuki Ebata, and others.
Research Institute for Languages and Cultures of Asia and Africa, Tokyo University of Foreign Studies.

UD Turkic.

Sep 2023 – Present
In collaboration with Furkan Akkurt, Dr. Çağrı Çöltekin, Dr. Jonathan Washington, and others.
Repository: <https://github.com/ud-turkic>

Language Documentation with an AI Helper.

Sep 2022 – Present
NSF Grant project in collaboration with Dr. Antonios Anastasopoulos, Dr. David Chiang, and Dr. Géraldine Walther.

Japanese SNACS (Semantic Network of Adposition and Case Supersenses).

Apr 2023 – Present
In collaboration with Tatsuya Aoyama, Dr. Nathan Schneider, and others (Georgetown University).

Building a multilingual speech-to-IPA system.

Jan 2023
SLT-Code Hackathon, IEEE.

Universal Dependencies: Tatar NMCTT treebank.

Aug 2021 – Jun 2022

Repository: <https://github.com/UniversalDependencies/UD.Tatar-NMCTT>

Talks

“Kichwa Meets Language Technologies: Automatic Speech Recognition for Kichwa”. *The Quechua Alliance 2024 Annual meeting*. Northwestern University, Chicago, IL, USA. October 13, 2024.

“El conjunto de datos de reconocimiento automático de voz para kichwa”. El *Taller: Tecnologías Digitales y Lenguas Indígenas*. Universidad de Chile, Santiago de Chile, Chile (presented online). January 10, 2024.

“Non-Discourse-Configurationality in Imbabura Kichwa”. *Linguistic Society of America Annual Meeting*. New York City, NY, USA. January 6, 2024.

“Building an Automatic Speech Recognition Dataset and Model for an Extremely Low-Resource Language”. *Notre Dame NL+*. University of Notre Dame, IN, USA. December 4, 2023.

“Reconocimiento automático del habla de kichwa para su inclusión y empoderamiento en el ciberespacio”. *VI Seminario Internacional Revitalizando Ando*. Universidad Politécnica Salesiana, Quito, Ecuador. November 21, 2023.

“UD-Tatar NMCTT Treebank: Issues in annotation across Turkic UD”. *UD Turkic Workshop 2023*. Istanbul Technical University, Istanbul, Turkey (presented online). September 8, 2023.

“Grammaticalization of modal nominal predicates in Tatar”. *International Conference on Turkish Linguistics*. Mainz, Germany. August 4, 2023.

“Bridging Natural Language Processing and Descriptive Linguistics with Universal Dependencies”. *Field Linguistics Workshop: Grammatical Studies Workshops 23 “Grammatical Studies and Digital Humanities (2)”*. Tokyo University of Foreign Studies, Tokyo, Japan (presented online). July 26, 2023.

“Japanese gapless relativization: The syntax–prosody interface to semantics”. *LFG23: the 28th International Lexical Functional Grammar Conference*. Rochester, NY, USA. July 24, 2023.

“Incorporating AI-based speech transcription into language documentation: A case study of Imbabura Kichwa”. *LSA Institute*, Amherst, MA, USA. July 12, 2023.

“Cross-linguistic analysis of the Mermaid Constructions in LFG”. *33rd South of England LFG Meetings*. London, UK (presented online). October 29, 2022.

“Universal Dependencies Tatar for Code-Switching Detection”. *Notre Dame NL+*. University of Notre Dame, IN, USA. September 19, 2022.

“Mermaid Construction in Lexical Functional Grammar”. *LFG22: the 27th International Lexical Functional Grammar Conference*. University of Groningen, Groningen, the Netherlands. July 14, 2022.

“Consistent Grammatical Annotation of Turkic Languages for more universal Universal Dependencies”. *Workshop on Computational Linguistics on East Asian languages*. 29th International Conference on Head-Driven Phrase Structure Grammar, online. July 31, 2022.

“Mermaid Construction as raising with a nominal predicate”. *The 74th Language Lunch At Edinburgh*. University of Edinburgh, Edinburgh, UK (presented online). March 8, 2022.

“Text Processing of Open-Access Jinghpaw Data on Google Colaboratory”. *Follow-up meeting on Intensive Language Course 2019 (Jinghpaw) / The 3rd meeting on Kachin studies*. ILCAA, Tokyo University of Foreign Studies, Tokyo, Japan (presented online). October 16, 2021.

Awards

Conference Travel Grant. Jun 2024
Kellogg Institute for International Studies, University of Notre Dame.

Leadership Advancing Socially Engaged Research (LASER) Award. May 2024 – August 2025
University of Notre Dame.

Doctoral Student Research Grant. April 2024 – June 2025
Kellogg Institute for International Studies, University of Notre Dame.

Conference Presentation Grant. Jun 2024
Graduate Student Government, University of Notre Dame.

Conference Travel Grant. Jun 2023
Kellogg Institute for International Studies, University of Notre Dame.

Professionalization Grant. May 2023
Kellogg Institute for International Studies, University of Notre Dame.

Committee’s Honorable Mention (委員特別賞). Mar 2023
The 29th Annual Meeting of the Association for Natural Language Processing.

Conference Presentation Grant. Mar 2023
Graduate Student Government, University of Notre Dame.

Summer Language Abroad Grant. Mar 2023
The Center for the Study of Languages and Cultures, University of Notre Dame.

JASSO Scholarship for Graduate-level Study Abroad. Sep 2021 – Aug 2022
Ministry of Education, Culture, Sports, Science and Technology, Japan.

CICP NAIST Multilingual Corpus. Jun 2021 – Mar 2022
Nara Institute of Science and Technology.

Grand Prix Winner, Apr 2019
7th International Olympiad of the Tatar Language and Literature.
Ministry of Education and Science, Republic of Tatarstan, Russia.

Exchange Program Scholarship. Sep 2018 – Jun 2019
Gyomu Super Japan Dream Foundation.

Honorable Mention, Keio University Academic Writing Contest. Nov 2014

Keio University.

Skills

Programming: Python, R, C, \LaTeX , HTML/CSS.

Softwares and libraries: PyTorch, Tensorflow, Wandb, Git, AWS, Docker, ELAN, Praat.

Languages:

Native: Japanese

Fluent: English

Advanced: Russian, Tatar, Spanish

Intermediate: French, German, Korean, Mandarin Chinese, Kichwa

Coursework: Eastern Armenian, Palestinian Arabic, Latin, Ancient Greek, Dutch, Kazakh, Finnish, Kurmanji Kurdish, Japanese Sign Language, Indonesian, Burmese, Javanese, Thai, Catalan, Italian, Brazilian Portuguese, Taiwanese Min-Nan, Akkadian, Middle Egyptian, Cantonese, Turkish, Cambodian, Vietnamese, Persian, Sanskrit, Wolaytta, Modern Tibetan, Turkmen, Old Church Slavonic, Old Armenian, Somali, Zulu, Punjabi, Yoruba, Modern Hebrew, Lithuanian, Jinghpaw, Pashto, Kreyòl Ayisyen (Haitian Creole)

Fieldwork: Chaga–Rombo (Bantu), Ikema–Miyakoan (Ryukyuan), Ekagi (Papuan), Taizhou Wu (Sinitic)

Affiliations

Kellogg Institute Doctoral Student Affiliate

Association of Computational Linguistics (ACL)

Linguistic Society of America (LSA)

言語処理学会 [Association for Natural Language Processing]

日本語学会 [Linguistic Society of Japan]