Shaoyun Yu, PhD

E-mail: syu.id@outlook.com

Education

PhD	**Nagoya University**, Japan	March 2021
	Graduate School of Humanities	
	Thesis: Nominative-accusative asymmetry in the processing of case drop sentences	
	Advisor: Prof Katsuo Tamaoka	
MLitt	**Nagoya University**, Japan	March 2017
	Graduate School of Languages and Cultures	
	Thesis: The effects of case drop in the processing of Japanese active sentences	
	Advisor: Prof Katsuo Tamaoka	
Research Student	**Nagoya University**, Japan	April 2014
	Graduate School of Languages and Cultures	—March 2015
BEng	**Nanjing University of Science and Technology**, China	June 2012
	School of Computer Science and Engineering	
	Thesis: An ASP.NET-based customer relationship management (CRM) system	

Research Positions

Postdoc	**The Hong Kong Polytechnic University**, Hong Kong SAR	June 2021
	Department of Chinese and Bilingual Studies	—present
	PI: Prof Ping Li	
	Projects: (1) The alignment between human brains and large language models (LLMs) in language representation and processing; (2) Technologies, data, and learning: Digital language learning (DLL) and second language representation	
Research Assistant	**Nagoya University**, Japan	May 2020
	Graduate School of Humanities	—March 2021
	PI: Prof Katsuo Tamaoka	
	Responsibilities: Development of a digital handwriting experiment for psycholinguistic research on learners of Japanese kanji (Chinese characters)	
Research Assistant	**Nagoya University**, Japan	May 2019
	Graduate School of Humanities	—March 2020
	PI: Prof Katsuo Tamaoka	
	Responsibilities: Development of a touch screen-based cognitive experiment for the diagnosis of Japanese aphasia patients	

Research Interests

- **Deep language models and the alignment between AI and human brains**: Connecting LLMs and neuroimaging research to understand principles of language learning, representation, and processing in artificial neural networks and humans.

- **Visuospatial cognition, embodiment, and digital language learning**: Investigating contributions of visuospatial cognition and multisensory integration in language learning with virtual reality (VR) simulation.

- **Language processing in typologically diverse languages**: Mechanisms underlying word order and case marking in Chinese (SVO), Japanese, Korean (SOV, nominative-accusative alignment), and Tongan (VSO, ergative-absolutive alignment).

Skills

- **Computational language models**: Expertise in building, training, and leveraging deep learning language models (e.g., BERT and GPT-based transformer models) for research in language sciences.
- **Scientific programming**: Deep understanding of the Python, MATLAB, and R languages. Expertise in popular packages and toolkits including transformers, Nilearn, SPM, EEGLAB, tidyverse, etc. Expertise in using high performance computing (HPC) clusters.
- **General programming**: Experience in web and Qt application development using Python and JavaScript. Experience in Linux system management.
- **Statistical analysis:** Hierarchical Bayesian modelling using the Stan language; various statistical procedures using Python and R.
- **Neural and behavioral experiments**: fMRI (design and data analysis); EEG, Eye-tracking (design, conduct, and data analysis). Experiment design using Psychopy, E-Prime, and Experiment Builder.
- **Languages**: Chinese (native), Japanese (fluent), English (fluent).

Awards

- **Japanese Government (MEXT) Scholarship**, Super Global University Category
  April 2019—March 2020
- **Japanese Government (MEXT) Scholarship**, Super Global University Category
  April 2016—September 2016

Five Most Representative Research Outputs

1. **Yu, S.**, Gu, C., Huang, K., & Li, P. (2024). Predicting the next sentence (not word) in large language models: What model-brain alignment tells us about discourse comprehension. *Science Advances*, *10*(21). https://doi.org/10.1126/sciadv.adn7744 (**IF**: 11.7)
2. Tamaoka, K., **Yu, S.**, Zhang, J., Otsuka, Y., Lim, H., Koizumi, M., & Verdonschot, R. G. (2024). Syntactic structures in motion: Investigating word order variations in verb-final (Korean) and verb-initial (Tongan) languages. *Frontiers in Psychology*, *15*, 1360191. https://doi.org/10.3389/fpsyg.2024.1360191 (**IF**: 2.6)
3. Kajiura, M., Jeong, H., Kawata, N., **Yu, S.**, Kinoshita, T., Kawashima, R., & Sugiura, M. (2021). Brain activity predicts future learning success in intensive second language listening training. *Brain and Language*, *212*. https://doi.org/10.1016/j.bandl.2020.104839 (**IF**: 2.1)
4. **Yu, S.**, & Tamaoka, K. (2020). Trade-off effect in the processing of Korean case-drop sentences: An eye tracking investigation. *Human Behaviour and Brain, 1*(2)*,* 49-57. https://doi.org/10.37716/HBAB.2020010203
5. Kwon, N., & **Yu, S.** (2018). Experimental evidence for the productivity of total reduplication in Japanese ideophones and ordinary vocabulary. *Language Sciences*, *66*, 166–182. https://doi.org/10.1016/j.langsci.2017.09.005 (**IF**: 1.3)

Publication List

Journal Articles (peer-reviewed)

1. **Yu, S.**, Gu, C., Huang, K., & Li, P. (2024). Predicting the next sentence (not word) in large language models: What model-brain alignment tells us about discourse comprehension. *Science Advances*, *10*(21). https://doi.org/10.1126/sciadv.adn7744 (**IF**: 11.7)

2. Tamaoka, K., **Yu, S.**, Zhang, J., Otsuka, Y., Lim, H., Koizumi, M., & Verdonschot, R. G. (2024). Syntactic structures in motion: Investigating word order variations in verb-final (Korean) and verb-initial (Tongan) languages. *Frontiers in Psychology*, *15*, 1360191. https://doi.org/10.3389/fpsyg.2024.1360191 (**IF**: 2.6)

3. Tamaoka, K., **Yu, S.**, Zhang, J., Miwa, K., & Koizumi, M. (2024). Constructing a web-accessible lexical database for core Tongan vocabulary. *Folia Linguistica*. https://doi.org/10.1515/flin-2024-2053 (**IF**: 0.6)

4. Kajiura, M., Jeong, H., Kawata, N., **Yu, S.**, Kinoshita, T., Kawashima, R., & Sugiura, M. (2021). Brain activity predicts future learning success in intensive second language listening training. *Brain and Language*, *212*. https://doi.org/10.1016/j.bandl.2020.104839 (**IF**: 2.1, **Contribution**: development of the experimental program; analysis of the fMRI data)

5. **Yu, S.**, & Tamaoka, K. (2020). Trade-off effect in the processing of Korean case-drop sentences: An eye tracking investigation. *Human Behaviour and Brain, 1*(2)*, 49-57.* https://doi.org/10.37716/HBAB.2020010203

6. Kwon, N., & **Yu, S.** (2018). Experimental evidence for the productivity of total reduplication in Japanese ideophones and ordinary vocabulary. *Language Sciences*, *66*, 166–182. https://doi.org/10.1016/j.langsci.2017.09.005 (**IF**: 1.3)

7. **Yu, S.**, & Tamaoka, K. (2018). Age-related differences in the acceptability of non-canonical word orders in Mandarin Chinese. *Lingua Sinica, 4*. https://doi.org/10.1186/s40655-018-0035-x

8. Hayakawa, K., **Yu, S.**, Chu, X., & Tamaoka, K. (2017). Objective indexes of phonological similarity between Japanese and Chinese two-kanji compound words: A comparative study on objective and subjective indexes [in Japanese]. *Center for Japanese Language Education Kwansei Gakuin University, 6*, 21−34. http://hdl.handle.net/10236/00025910

Journal Articles (under review)

9. Kajiura, M., Jeong, H., Kawata, N., **Yu, S.**, Kinoshita, T., Kawashima, R., & Sugiura, M. (under review). Working memory as the best predictor of future Learning success for L2 listening: the impact of varying processing demands.

Journal Articles (in preparation)

10. Xu, Z., **Yu, S.**, Cai, Z. (in preparation). Using OpenHandWrite with PsychoPy Builder for capturing handwriting processes: A stroke-level database of Chinese character handwriting.

Journal Articles (non peer-reviewed)

11. **Yu, S.**, & Tamaoka, K. (2019). Construction of a database and search engine on the phonological similarity of two-kanji compound words in Japanese, Korean, Chinese and Vietnamese [in Japanese]. *Studia Linguistica*, *33*, 75-94. Nagoya University. https://doi.org/10.18999/stul.33.75

12. Hoang, T.L.P., Tamaoka, K., **& Yu, S.** (2019). Developing objective indexes of phonological similarities for two-kanji compound words used in both the Japanese and

Vietnamese languages [in Japanese]. *Studia Linguistica*, *33*, 133-146. Nagoya University. https://doi.org/10.18999/stul.33.133 (**Contribution**: computation of the indexes; interpretation of the results)

13. **Yu, S.**, & Tamaoka, K. (2015). A web-accessible search engine for grammatical category of orthographically-similar two-kanji compound words between Japanese, Korean and Chinese [in Japanese]. *Studia Linguistica*, *29*, 43-61. Nagoya University. https://doi.org/10.18999/stul.29.43

Book chapters

1. **Yu, S.**, Gao, Z., Zhao, J., Yu, S., & Li, P., (under review). Immersive virtual reality for second language learning.

2. Gao, Z., **Yu, S.**, Zhao, J., & Li, P. (under review). Virtual reality (VR) for language learning: Methods and approaches. In A. Witt, A. Lenz, & P. Kamocki (Eds.), *Routledge Handbook of Digital Linguistics*. Routledge.

Conference Presentations (peer-reviewed)

1. **Yu, S.**, Gu, C., Huang, K., & Li, P. (2024, April). *Predicting the next sentence (not word) in pretraining: What model-brain alignment tells us about discourse-level comprehension*. The Cognitive Neuroscience Society (CNS) 2024 Annual Meeting, Toronto, Canada.

2. **Yu, S.**, Gu, C., Huang, K., & Li, P. (2023, March). *Representations from deep language models capture neural patterns in naturalistic reading of scientific texts: cognitive plausibility and neural relevance*. The Cognitive Neuroscience Society (CNS) 30th Anniversary Meeting, San Francisco, USA.

3. Tamaoka, K., **Yu, S.**, Zhang, J., & Koizumi, M. (2022, March). *Observing the topicalization effect in Tongan sentence processing*. The 35th Annual Conference on Human Sentence Processing (HSP2022), Online, UC Santa Cruz, USA.

4. **Yu, S.**, & Tamaoka, K. (2019, March). *The online processing of case-drop sentences in Korean: An eye tracking investigation*. The 2019 Annual Meeting of the International Society of Neuroscience (ISON), Shenzhen Convention and Exhibition Center, China.

5. Hoang, T.L.P., Tamaoka, K., **& Yu, S.** (2019, May). *The effects of phonological similarity and usage frequency on the translation of kanji words by Vietnamese learners of Japanese* [ベトナム人日本語学習者による漢字語の翻訳における音韻類似性と使用頻度の影響]. The 2019 Spring Meeting of the Society for Teaching Japanese as a Foreign Language, International Congress Center Epochal Tsukuba, Japan.

6. **Yu, S.**, & Tamaoka, K. (2018, November). A contrastive study on the asymmetry of nominative and accusative case drops in Japanese and Korean. *Proceedings of the 157th Meeting of the Linguistic Society of Japan*, 300–305.

7. **Yu, S.**, Kim, J. & Tamaoka, K. (2018, May). *Construction of an online search engine for a database on phonological similarity and phonological distance of two-kanji compound words in Japanese, Korean, Chinese and Vietnamese* [２字漢字語の音韻類似性・音韻的距離に関する日韓中越データベースのオンライン検索エンジンの構築]. The 2018 Spring Meeting of the Society for Teaching Japanese as a Foreign Language, Tokyo University of Foreign Studies, Japan.

8. Kajiura, M., Jeong, H., Kawata, N., **Yu, S.**, Kinoshita, T., Kawashima, R., & Sugiura, M. (2018, August). *Brain activity predicts future learning potential in intensive second language listening training.* The Tenth Annual Meeting of the Society for the Neurobiology of Language, Québec City Convention Centre, Canada.

9. Kajiura, M., Jeong, H., Kawata, N., **Yu, S.**, Kinoshita, T., Kawashima, R., & Sugiura, M. (2018, June). *The effects of L2 fast-rate listening training combined with transcript reading*

and brain activity. The 24th Annual Meeting of the Organization for Human Brain Mapping (OHBM), Suntec Singapore International Convention and Exhibition Centre, Singapore.

10. Kajiura, M., **Yu, S.**, & Kinoshita, T. (2018, August). *Working memory capacity as the best predictor of future learning potential for L2 listening*. The 58th National Conference of The Japan Association for Language Education and Technology (LET), Senri Life Science Center, Japan.

11. Kajiura, M., **Yu, S.**, Jeong, H., Kinoshita, T., & Oishi, H. (2017, August). *Investigating the mapping between phonological and semantic representations during L2 listening: A comparison based on L1-L2 differences and proficiency* [L2 聴解時の音声と意味表象のマッピング処理に関する検証 −表象言語（L1・L2）の違いと熟達度による比較−]. The 43th Conference of the Japan Society of English Language Education. Shimane University, Japan.

12. Tamaoka, K., **Yu, S.**, & Mansbridge, M. (2015, September). *An eye-tracking investigation of the processing of nominative/genitive-alternations in Japanese*. Architectures & Mechanisms for Language Processing 2015, University of Malta Valletta Campus, Malta.

13. **Yu, S.**, Xiong, K., Hayakawa, K., & Tamaoka, K. (2015, October). *Construction of an online search engine for a database on grammatical categories of two-kanji compound words in Japanese, Korean, and Chinese* [同形二字漢字語の品詞性に関する日韓中データベースのオンライン検索エンジンの構築]. The 2018 Fall Meeting of the Society for Teaching Japanese as a Foreign Language, Okinawa International University, Japan.

Computer Programs

1. **Yu, S.** (2024). The Tongan Lexicon Project: Web database for commonly-used 4,013 Tongan words. https://tonganlex.vercel.app/

2. **Yu, S.** (2018). KanjigoDB: A database on two-kanji compound words in Japanese, Korean, Chinese, and Vietnamese. http://kanjigodb.herokuapp.com

3. **Yu, S.** (2016a). An alternative implementation of the MATTR algorithm in R. https://github.com/rongmu/mattr

4. **Yu, S.** (2016b). phonosim: An experimental R package for calculating phonological similarity. https://github.com/rongmu/phonosim

5. **Yu, S.** (2015). BunCuts: A text formatter for Japanese. https://github.com/rongmu/buncuts

Workshops / Talks

1. **Yu, S.** (2016, March). *Corpus data processing based on data frames* [データフレームに基づいたコーパスの処理]. Twenty-minute talk at Nagoya.R #15, Nagoya University.

2. **Yu, S.** (2015, October). *Practising the Tidy Data framework: Processing survey data acquired with SDAPS* [Tidy Data フレームワークの実践：SDAPS の質問紙データの処理]. Twenty-minute talk at Nagoya.R #14, Nagoya University.

3. **Yu, S.** (2015, February). *Corpus database management and search using ChaKi.NET* [ChaKi.NET を使ったコーパス管理と検索]. Three-hour workshop at Nagoya University, organized by the Japanese Society for Language Sciences.

Professional Service

- **Peer Review**
  Journals*: Cognitive Science*, *Brain and Language*
  Conferences: *Annual Meeting of the Cognitive Science Society*