Japanese Adverb Taxonomy: Modern NLP Tools and Comparative Linguistic Analysis

Eric Odle[1], Yun-Ju Hsueh[2], Soufiane Ghzal, and Pei-Chun Lin[3]
[1]Hokkaido University (Japan) [2]Yuan Ze University (Taiwan) [3]Feng Chia University (Taiwan)
[1]ericmichael.odle.q5@elms.hokudai.ac.jp
[2]conniehsueh.1984@gmail.com
[3]peiclin@mail.fcu.edu.tw

Abstract

This paper compares Japanese adverb taxonomies with results from natural language processing analysis. A list of 3,801 Japanese adverbs was generated using the Japanese Multilingual Dictionary (JMDICT). Next, word embeddings were produced from these adverbs using four chiVe (sudachi Vectors) models. Embeddings were then clustered using k-means and hierarchical clustering algorithms, comparing results with the Yamada 3-category and Noda 5-category linguistic taxonomies. Additionally, Silhouette analysis indicated optimal clustering at k=3 and k=5 clusters. Manual labeling of a random adverb subset showed that the Yamada taxonomy tends to unevenly represent adverbs: 66.3% Status vs. 14.2% Degree and 19.5% Declarative. However, the Yamada taxonomy achieved a higher classification agreement (62.7%) with embedding clusters than the Noda taxonomy. Overall, this research contributes insight into Japanese adverb categorization and sets the stage for future studies.

1 Introduction

Taxonomic classification of Japanese adverbs is a longstanding aim among Japanese linguists [1] and second language educators [2]. Multiple researchers have attempted to systematize Japanese adverbs through categorical taxonomies [3, 4, 5, 6, 7, 8], but differ by linguistic theory and number of categories. Despite numerous attempts, Yamada's 3-category (Status, Degree, and Declarative) taxonomy from 1936 [9] remains one of the oldest and most commonly used systems for systematizing Japanese adverbs.

Many tools are currently available to aid this pursuit. For example, the Japanese Multilingual Dictionary (JMDICT) [10] serves as a reference for various applications, including second language education [11, 12, 13] and natural language processing (NLP) research [14, 15]. Word2Vec [16] is another common NLP technique, utilizing a simple neural network to represent words as multi-dimensional vectors. Word2Vec's impact on NLP includes providing efficient word embeddings for machine translation [17] and other language modeling tasks [18, 19]. This success has led to the development of other word embedding techniques, such as Sudachi [20] and chiVe (sudachi Vectors) [21]. Sudachi, an open-source Japanese morphological analyzer, breaks down Japanese text into morphemes, aiding in tasks like text analysis and machine translation. Consequently, these tools have led to increased interest [22, 23, 24, 25] in analyzing the relationship between semantics and embeddings.

2 Methodology

Here, the Japanese Multilingual Dictionary (JMDICT) was used to generate a comprehensive list of Japanese adverbs. After list assembly, a normalization step was performed during which the first listed kanji (Chinese script) form of the adverb was favored. If no kanji form was present, the kana (Japanese script) form was used. The final list contained 3,801 adverbs.

To examine semantic relationships among adverbs, we utilized four chiVe models to generate embeddings: chiVe-1.1-mc5, chiVe-1.1-mc90, chiVe-1.2-mc5, and chiVe-1.2-mc90. Embeddings were clustered using either the k-means algorithm [26] or agglomerative hierarchical clustering [27, 28, 29] driven by the k-means algorithm. The Python library scikit-learn [30] was used to cluster adverb

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

embeddings. Clusters were labeled arbitrarily in numerical order (starting with Cluster 0). Silhouette scores [31] were generated for each embedding set to determine the optimal number of clusters per test. The t-Distributed Stochastic Neighbor Embedding (t-SNE) [32] method was chosen over Principal Component Analysis (PCA) [33] for plotting high-dimensional word embeddings in 2D. This selection was based on the effectiveness of t-SNE in capturing local relationships and intricate patterns.

When comparing embedding clusters to conventional linguistic taxonomies, we extracted a random subset of 395 adverbs from the JMDICT dataset. This subset was manually labeled based on the Yamada 3-category and Noda 5-category taxonomies. To evaluate agreement between linguistic taxonomies and cluster assignments, we conducted a rigorous comparative analysis. Yamada's 3-category taxonomy was compared to k=3 cluster results, while Noda's 5-category taxonomy was compared to k=5 cluster results. This involved comparing each human-assigned category with each cluster, then selecting the permutation with the highest agreement.

3 Results

3.1 Taxonomic representation among Japanese adverbs

Manual labeling of a random adverb subset (395 adverbs) using the Yamada and Noda systems resulted in unequal representation among both taxonomies (Table 1).

Table 1 Taxonomic representation across a Japanese adverb subset

Taxonomy	Representation (%)				
	Status	Degree	Declar.		
Yamada	66.3	14.2	19.5		
	Mood	Tense	Aspect	Voice	Object
Noda	14.9	12.7	22.3	14.7	35.4

Under the Yamada taxonomy, Status adverbs were observed for the majority (66.3%) of adverbs compared to the less frequent Degree (14.2%) and Declarative (19.5%). Under the Noda taxonomy, category representation was more uniform, showing a slight bias towards Aspect (22.3%) and Object (35.4) adverbs.

3.2 Both k=3 and k=5 fit the data well

Silhouette testing was performed on JMDICT set (3,801 adverbs) embeddings generated using four chiVe models (Figure 1). Adverb embeddings clustered using the k-means algorithm optimally at k=3 and k=5 all for chiVe models tested (k=2 excluded).

3.3 Cluster visualization reveals model-dependent patterns

Adverb clusters generated from various chiVe model and clustering algorithm combinations were visualized as 2D plots (t-SNE), showing a consistent pattern of one large mass and one smaller island of points. For k=3 clusters (Figure 2), both k-means and hierarchical clustering methods formed two clusters sharing a fuzzy boundary within the large mass. Similarly, k=5 clusters (Figure 3) resulted in a central mass of four clusters sharing fuzzy boundaries and a smaller cluster island.

3.4 Embedding cluster agreement varies with taxonomy

Agreement between cluster assignment and adverb category was higher with the Yamada taxonomy than the Noda taxonomy (Table 2). The highest accuracy comparing k=3 clusters to the Yamada taxonomy was 0.627, while the highest accuracy comparing k=5 clusters to the Noda taxonomy was 0.406.

Table 2 Classification agreement between embedding clusters and linguistic taxonomies

Model	Yamada	Noda
k-means clustering		
1.1-mc5	0.508	0.368
1.1-mc90	0.546	0.368
1.2-mc5	0.485	0.322
1.2-mc90	0.536	0.325
hierarchical clustering		
1.1-mc5	**0.627**	0.368
1.1-mc90	0.470	**0.406**
1.2-mc5	0.546	0.350
1.2-mc90	0.437	0.383

Top accuracies for both Yamada and Noda taxonomies were achieved using version 1.1 chiVe models and hierarchical clustering. For the Yamada taxonomy, chiVe 1.1-mc5 outperformed the second highest accuracy (0.546) by 8.1%. For the Noda taxonomy, chiVe 1.1-mc90 outperformed the second highest accuracy (0.383) by 2.3%.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

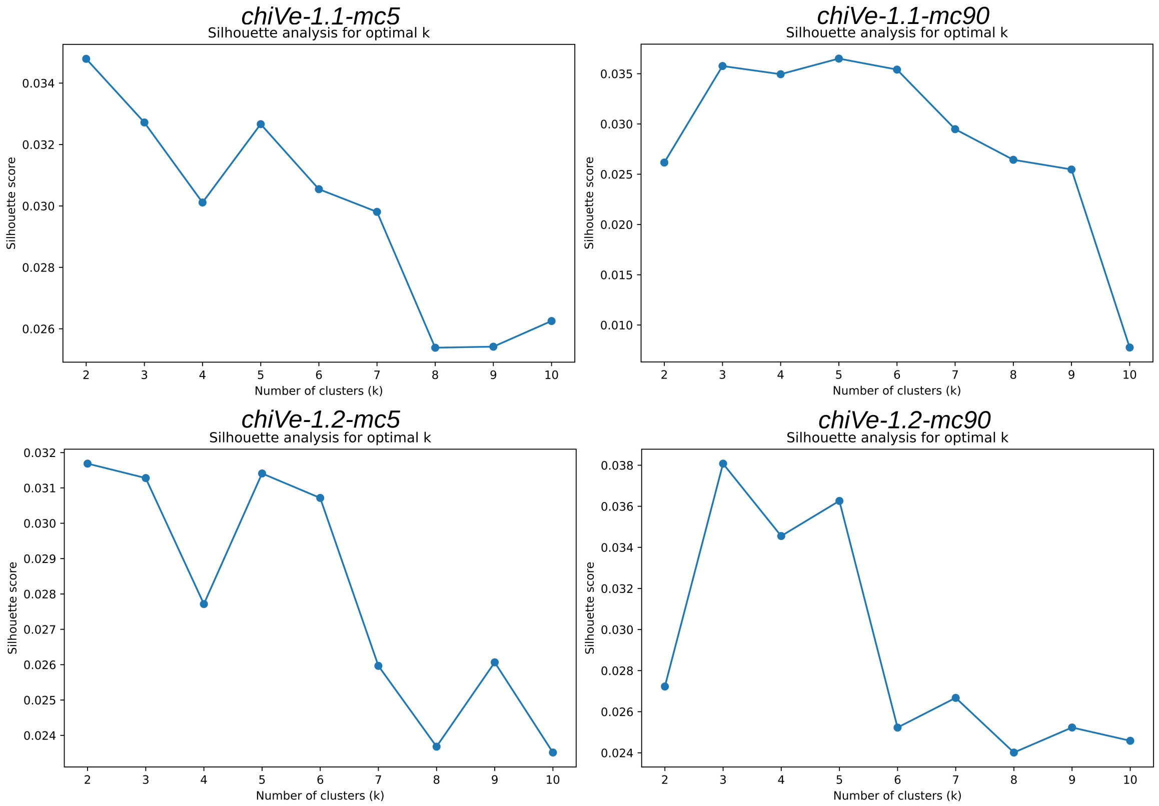

Figure 1 Silhouette analysis plots for adverb embeddings generated using four chiVe models.

4 Discussion

4.1 Taxonomic heterogeneity among Japanese adverbs

Unequal representation from the Yamada taxonomy (Table 11) suggested that some categories may be overly general. A common assumption is that optimal taxonomies should be represented evenly: i.e., if the Yamada taxonomy is a good system, then there should be an equal number of Status, Degree, and Declarative adverbs in the Japanese language. However, there is no theoretical necessity for this. Our results (Table 2) show a consistently higher agreement between embedding clusters and Yamada categories, suggesting an inherent imbalance of adverb types in Japanese.

4.2 Hierarchical structure among Japanese adverb embeddings

Silhouette testing was performed to quantitatively approach the question of optimal category number, and results showed maxima at k=3 and k=5 for all four models tested (Figure 1). High Silhouette scores at k=2 suggest a latent hierarchical structure among Japanese adverbs. Moreover, visualization of both k=3 (Figure 2) and k=5 (Figure 3) tests suggest that subdivision of the large, central mass of points is likely influenced by model parameters. Finally, the highest agreement scores were achieved using hierarchical clustering: 62.7% vs. 48.3% reported previously [25]. Together, these findings suggest a hierarchical structure among Japanese adverbs.

4.3 Methodological challenges and future optimizations

This study highlights multiple areas for improvement. First, alternative taxonomies [3, 4, 6, 7, 8] based on other linguistic theories could point to better methods of Japanese adverb categorization. Second, future work would benefit from more trained researchers tasked with labeling adverbs. Reliability of the JMDICT-generated adverb list could also be improved by a multi-rater review, as JMDICT is not always stringently curated [34]. Third, large-scale labeling of all 3,801 JMDICT adverbs would provide base rate estimates, equipping researchers with the ability to generate balanced ground truth datasets. These datasets could then be used for training classifiers, fine-tuning large language models, and other NLP applications.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

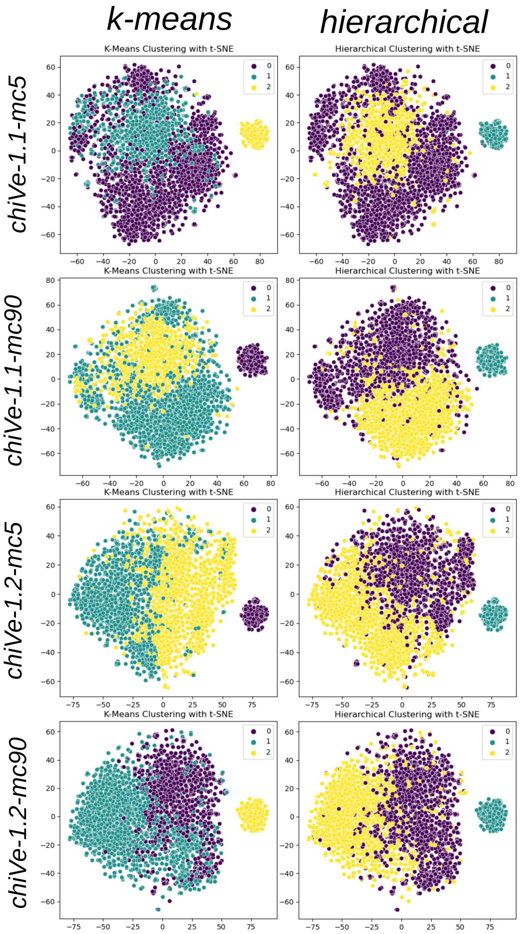

Figure 2 Japanese adverb clusters embedded using four chiVe models; k=3. Cluster membership (color) is determined by clustering algorithm (k-means vs. hierarchical), while positional in lower dimensional space is determined by dimensional reduction method (t-SNE). Membership labels are arbitrarily assigned.

5 Conclusion

In summary, this study explored the taxonomy of Japanese adverbs through chiVe model embeddings, challenging conventional assumptions and revealing a nuanced hierarchical structure. Comparison between Yamada and Noda taxonomies highlights possible unequal representation, prompting a re-evaluation of base assumptions about the Japanese language. Silhouette analysis indicated a latent hierarchy with peaks at k=3 and k=5. Moreover, hierarchical clustering yielded superior agreement with conventional classification taxonomies. Overall, the present study contributes to the fields of Japanese linguistics and semantic analysis, providing a foundation for future studies on these topics.

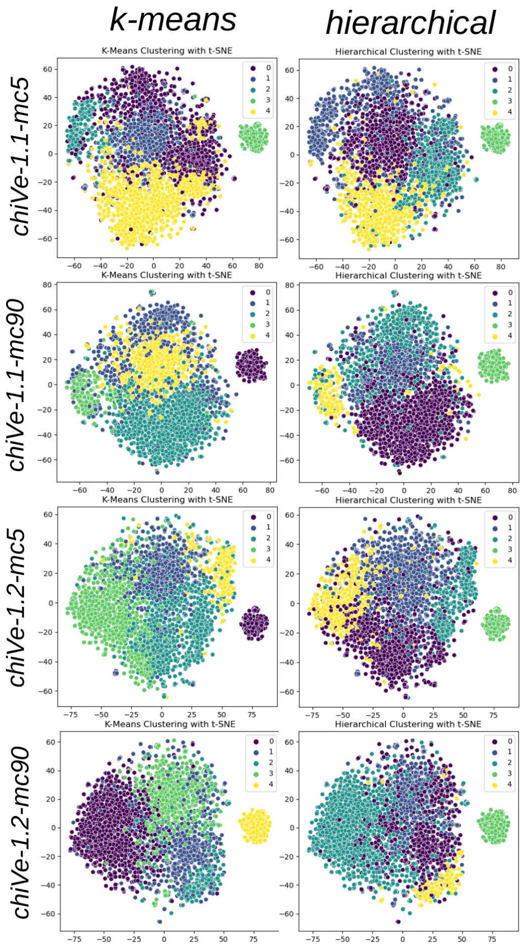

Figure 3 Japanese adverb clusters embedded using four chiVe models; k=5. Cluster membership (color) is determined by clustering algorithm (k-means vs. hierarchical), while positional in lower dimensional space is determined by dimensional reduction method (t-SNE). Membership labels are arbitrarily assigned.

5.1 Supplemental Material

Please refer to the project GitHub repository for supplemental material.

References

[1] Koichiro Nakamura. Towards a unified analysis of japanese adverb types and their syntactic positions. **Florida Linguistics Papers**, Vol. 4, No. 4, 2017.

[2] Masako HAYASHI. **A study of adverbs presented in textbooks for Japanese learners**. PhD thesis, Tohoku University, 2018.

[3] Minoru Nakau. **Ninchi imiron no genri (Principles of cognitive semantics)**. Taishukan, Tokyo, 1994.

[4] Toshiyuki Kanamaru, Masaki Murata, and Hitoshi Isahara. Creation of a japanese adverb dictionary that includes information on the speaker's communicative intention using machine learning. In **LREC**, pp. 1706–1709, 2006.

[5] Hisashi Noda. Fukusi-no gojyun. **Nihongo Kyooiku**, Vol. 52, pp. 79–90, 1984.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

[6] Yoshio Endo. **Locality and Information Structure**. John Benjamins Publishing Company, Amsterdam, 2007.

[7] Kentaro Ogura, Francis Bond, and Satoru Ikehara. A method of ordering english adverbs. journal of natural language processing. **Journal of Machine Learning Research**, Vol. 4, pp. 17–39, 1997.

[8] Kazuma Fujimaki. On the relative structural position of high adverbs and the interpretation of ga-marked subject. In **139th meeting of the linguistic society of Japan**, 2009.

[9] Yoshio Yamada. **Nihon Bunpou Gaku Gairon (Survey of Japanese Grammar)**. Houbun Kan, Tokyo, 1936.

[10] Jim Breen. Jmdict: a japanese-multilingual dictionary. In **Proceedings of the workshop on multilingual linguistic resources**, pp. 65–72, 2004.

[11] Mario E Aburto-Gutierrez, Gamar Azuaje, Vipul Mishra, Shaira Osmani, and Kazushi Ikeda. Jasenpai: Towards an adaptive and social interactive e-learning platform for japanese language learning. In **2022 International Conference on Advanced Learning Technologies (ICALT)**, pp. 236–238. IEEE, 2022.

[12] Bartholomäus Wloka and Werner Winiwarter. Dare–a comprehensive methodology for mastering kanji. In **The 23rd International Conference on Information Integration and Web Intelligence**, pp. 425–433, 2021.

[13] Werner Winiwarter. Mastering japanese through augmented browsing. In **Proceedings of International Conference on Information Integration and Web-based Applications & Services**, pp. 179–188, 2013.

[14] Ghazal Afroozi Milani, Daniel Cyrus, and Alireza Tamaddoni-Nezhad. Towards one-shot learning for text classification using inductive logic programming. **arXiv preprint arXiv:2308.15885**, 2023.

[15] Francis Bond, Eric Nichols, and Jim Breen. Enhancing a dictionary for transfer rule acquisition. **Linguistic Research**, Vol. 24, No. 2, pp. 133–151, 2007.

[16] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. **arXiv preprint arXiv:1301.3781**, 2013.

[17] Shashi Pal Singh, Ajai Kumar, Hemant Darbari, Lenali Singh, Anshika Rastogi, and Shikha Jain. Machine translation using deep learning: An overview. In **2017 international conference on computer, communications and electronics (comptelix)**, pp. 162–167. IEEE, 2017.

[18] Kazunori Yawata, Tamao Suzuki, Keisuke Kiryu, and Ken Mohri. Performance evaluation of japanese bert model for intent classification using a chatbot. 人工知能学会全国大会論文集 第35回 (2021), pp. 2N4IS2c05–2N4IS2c05. 一般社団法人 人工知能学会, 2021.

[19] Ekaterina Popova and Vladimir Spitsyn. Sentiment analysis of short russian texts using bert and word2vec embeddings. In **Graphion conferences on computer graphics and vision**, Vol. 31, pp. 1011–1016, 2021.

[20] Kazuma Takaoka, Sorami Hisamoto, Noriko Kawahara, Miho Sakamoto, Yoshitaka Uchida, and Yuji Matsumoto. Sudachi: A japanese tokenizer for business. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**, 2018.

[21] Sorami Hisamoto, Takashi Yamamura, Akihiro Katsuta, Yuto Takebayashi, Kazuma Takaoka, Yoshitaka Uchida, Teruaki Oka, and Masayuki Asahara. chive: Towards industrial-strength japanese word vector resources–constructing and improving embedding with tokenizer. **IEICE Technical Report; IEICE Tech. Rep.**, Vol. 120, No. 166, pp. 40–45, 2020.

[22] Shusen Liu, Peer-Timo Bremer, Jayaraman J Thiagarajan, Vivek Srikumar, Bei Wang, Yarden Livnat, and Valerio Pascucci. Visual exploration of semantic relationships in neural word embeddings. **IEEE transactions on visualization and computer graphics**, Vol. 24, No. 1, pp. 553–562, 2017.

[23] Derry Jatnika, Moch Arif Bijaksana, and Arie Ardiyanti Suryani. Word2vec model analysis for semantic similarities in english words. **Procedia Computer Science**, Vol. 157, pp. 160–167, 2019.

[24] LV Savytska, NM Vnukova, IV Bezugla, Vasyl Pyvovarov, and M Turgut Sübay. Using word2vec technique to determine semantic and morphologic similarity in embedded words of the ukrainian language. 2021.

[25] Eric Odle, Yun-Ju Hsueh, and Pei-Chun Lin. Semantic positioning model incorporating bert/roberta and fuzzy theory achieves more nuanced japanese adverb clustering. **Electronics**, Vol. 12, No. 19, p. 4185, 2023.

[26] Stuart Lloyd. Least squares quantization in pcm. **IEEE transactions on information theory**, Vol. 28, No. 2, pp. 129–137, 1982.

[27] Cecil C Bridges Jr. Hierarchical cluster analysis. **Psychological reports**, Vol. 18, No. 3, pp. 851–854, 1966.

[28] Fionn Murtagh and Pedro Contreras. Algorithms for hierarchical clustering: an overview. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, Vol. 2, No. 1, pp. 86–97, 2012.

[29] Fionn Murtagh and Pedro Contreras. Algorithms for hierarchical clustering: an overview, ii. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, Vol. 7, No. 6, p. e1219, 2017.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, Vol. 12, pp. 2825–2830, 2011.

[31] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of computational and applied mathematics**, Vol. 20, pp. 53–65, 1987.

[32] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. **Journal of machine learning research**, Vol. 9, No. 11, 2008.

[33] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. **The London, Edinburgh, and Dublin philosophical magazine and journal of science**, Vol. 2, No. 11, pp. 559–572, 1901.

[34] Shigeki Karita, Richard Sproat, and Haruko Ishikawa. Lenient evaluation of japanese speech recognition: Modeling naturally occurring spelling inconsistency. **arXiv preprint arXiv:2306.04530**, 2023.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).