Long Unit Word Tokenization and Bunsetsu Segmentation of Historical Japanese

Long Unit Word Tokenization and Bunsetsu Segmentation of Historical Japanese

Hiroaki Ozaki[1] **Komiya Kanako**[1] **Masayuki Asahara**[2] and **Toshinobu Ogiso**[2]
[1]Tokyo University of Agriculture and Technology, Japan
[2]National Institute for Japanese Language and Linguistics, Japan
hiroaki-ozaki@st.go.tuat.ac.jp, kkomiya@go.tuat.ac.jp,
{masayu-a,togiso}@ninjal.ac.jp

Abstract

In Japanese, "bunsetsu" is the natural minimal phrase of a sentence; it serves as a natural boundary of a sentence for native speakers rather than words, and thus grammatical analysis in Japanese linguistics commonly operates on the basis of bunsetsu units. By contrast, because Japanese does not have delimiters between words, there are two major categories of word definitions: Short Unit Words (SUWs) and Long Unit Words (LUWs). SUW dictionaries are available, whereas LUW dictionaries are not. Hence, this study focuses on providing deep learning-based (or LLM-based) bunsetsu and LUWs parser for the Heian period (AD 794-1185) and evaluating its performances. We model the parser as a transformer-based joint sequential labels model that combines the bunsetsu BI tag, LUW BI tag, and LUW Part-of-Speech (POS) tag for each SUW token. We trained our models on the corpora of each period including contemporary and historical Japanese. The results ranged from 0.976 to 0.996 in the f1 value for both bunsetsu and LUW reconstruction indicating that our models achieved comparable performance with models for a contemporary Japanese corpus. Through statistical analysis and a diachronic case study, it was found that the estimation of bunsetsu could be influenced by the grammaticalization of morphemes.

1 Introduction

In Japanese, "bunsetsu" (base-phrase) is the natural minimal phrase of a sentence. It serves as a natural boundary of a sentence for native speakers rather than words; thus grammatical analysis in Japanese linguistics commonly operates on the basis of bunsetsu units. For example, in Universal Dependencies (UD; Nivre et al., 2020), a framework for the consistent annotation of lexical dependency grammar across different human languages, some Japanese corpora have been converted from dependency relations between bunsetsu (Asahara et al., 2018; Omura and Asahara, 2018).

In contrast, because Japanese does not have delimiters between words, there are many definitions of "words" in Japanese. The National Institute for Japanese Language and Linguistics defines two hierarchical word tokenization categories: Short Unit Words (SUWs) and Long Unit Words (LUWs). SUW is a minimal word token in Japanese, and is defined by a bottom-up method that consists of at most two morphological units. In contrast, LUW is defined by a top-down method that divides a bunsetsu into two parts, and it may contain several SUWs. For example, the LUW "北西大西洋 (Northwest Atlantic)" consists of two SUWs "北西 (Northwest)" and "大西洋 (Atlantic)."

Dictionaries of SUWs for historical and contemporary Japanese are already publicly available [1], whereas there is no dictionary for LUWs. Hence, a parser that outputs bunsetsu and LUWs for historical Japanese is necessary to analyze the grammatical changes in Japanese.

For existing historical Japanese literature, a sufficient amount of bunsetsu and LUW annotated text to train the parser is primarily available from the Heian period (AD 794-1185) and later. Therefore, this study mainly focuses on the Heian period, with the subsequent Kamakura (AD 1185-1336) and the Muromachi (AD 1336-1573) periods chosen for comparison.

The existing bunsetsu parser (Kozawa et al., 2014) for these periods is based on Conditional Random Field (CRF), which was used to create the annotated corpus. Thus, this study focuses on providing a deep learning-based (or LLM-based) bunsetsu and Long Unit Words (LUW) parser and evaluating its performances. We model the parser

[1] https://clrd.ninjal.ac.jp/unidic/

Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024), pages 48–55

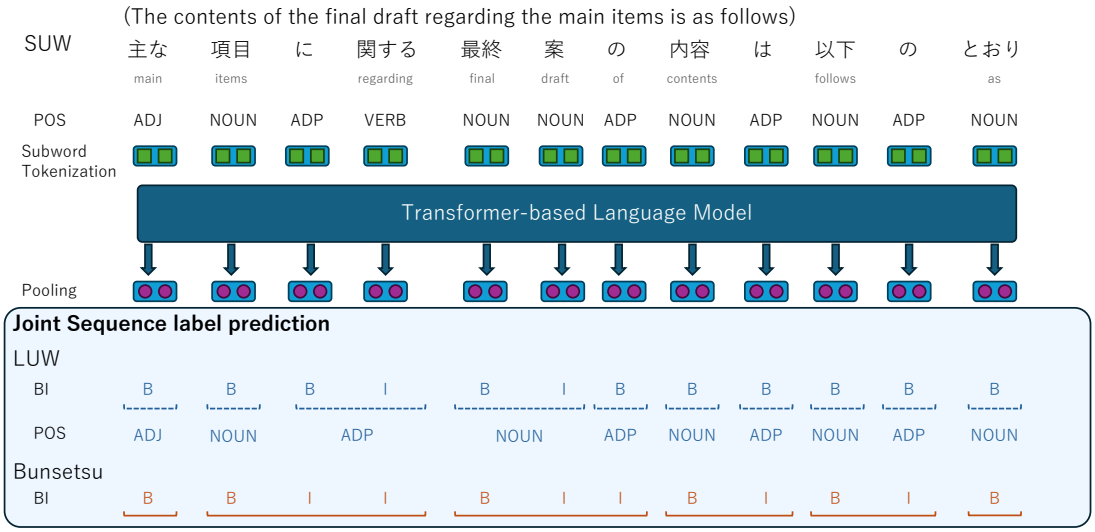


Figure 1: Overview of bunsetsu and Long Unit Words (LUWs) tokenization.

as a joint sequential label that combines the bunsetsu BI tag, LUW BI tag, and LUW Part-of-Speech (POS) tags for each SUW token. We used a Transformer-based Language Model (TLM) to output an SUW token representation by taking the appropriate pooling of subword representations for the last layer of the transformer. We preserved the SUW boundaries when tokenizing a given sentence into subwords. We trained our models on the corpora of each period including contemporary and historical Japanese.

The results indicate that the models trained on historical Japanese achieve comparable performance (0.976-0.996 f1 values) to a model for a contemporary Japanese corpus. To trace grammatical changes in Japanese, we evaluated the zero-shot transfer performance of the Heian, Kamakura, and Muromachi periods for each other. The models trained with a corpus of the Heian and Kamakura periods performed well on each other, whereas the model trained with a corpus of the Muromachi period did not. These results support the consensus among Japanese liguists that the large grammatical changes occurred during the Muromachi period. Furthermore, the analysis focusing on sentence-ending particles revealed that new sentence-ending particle usage has emerged in the Muromachi, and they are difficult to predict by the models of the prior periods.[2]

[2]Our code is publicly available at https://github.com/komiya-lab/monaka

2 Related Work

Parser for Historical Japanese Comainu, a Japanese bunsetsu and LUW parser, was originally provided for contemporary Japanese (Kozawa et al., 2014), although it can also be applied to historical Japanese. Comainu takes SUW tokens as input, which are tokenized by a CRF-based morphological analyzer MeCab[3], and then outputs the bunsetsu and LUW tokens. As mentioned above, Comainu is a CRF-based parser; thus, we focused on deep-learning-based methods.

Parser for Contemporary Japanese Recent Japanese corpora of UD contain bunsetsu and LUW annotations (Omura et al., 2023); thus, some parsers trained on these corpora support bunsetsu segmentation and LUW tokenization. For example, the spaCy-based [4] Japanese UD parser[5] supports LUW tokenization (Matsuda et al., 2022). The parser was trained with a Transformer-based language model (TLM) through the spaCy pipeline, and it achieved better performance than Comainu by adding some rules. GiNZA (Matsuda, 2020), which is also a spaCy-based parser, supports bunsetsu output.

3 Bunsetsu and Long Unit Word

3.1 Short Unit Word

Short Unit Word (SUW) is a token close to the granularity of typical Japanese word tokens. A

[3]https://taku910.github.io/mecab/
[4]https://spacy.io/
[5]https://github.com/megagonlabs/UD_Japanese-GSD/releases/tag/r2.9-NE/

	Heian	Kamakura	Muromachi	UD-Japanese-GSD
Number of				
Sentence	196,680	332,575	154,080	8,100
SUW	5,084,245	6,519,090	2,077,960	193,654
LUW	4,576,115	6,003,790	1,923,300	150,244
Bunsetsu	1,986,150	2,700,520	881,015	65,966
Average numbers in a sentence				
Characters	43.029	27.779	21.511	39.371
SUW	25.850	19.602	13.486	23.908
LUW	23.267	18.052	12.482	18.549
Bunsetsu	10.098	8.120	5.718	8.144

Table 1: Statistics of the Corpus of Historical Japanese (CHJ) (Heian, Kamakura, and Muromachi) and UD-Japanese-GSD.

	UD-Japanese-GSD	CHJ
dropout rate	0.5	0.5
dim. POS emb.	256	256
learning rate	2e-05	5e-06
batch size	28	24
num. of epocsh	50	20
gradient clip	5.0	5.0
gradient decay	0.75	0.75
decay step	5000	5000

Table 2: Hyperparamters

dictionary (UniDic) was established for SUWs, enabling high-performance morphological analysis based on UniDic (Den et al., 2008). As shown in the overview Figure 1, bunsetsu and LUWs are also composed of SUWs.

3.2 Bunsetsu (Base-pharase)

A bunsetsu is a (natural) minimal phrase that consists of a Japanese sentence. Generally, a bunsetsu boundary occurs after a particle or a sequence of particles. This is because Japanese functional words typically follow their content words, on which they depend. In Figure 1, all LUW noun (NOUN) and adposition (ADP) pairs are composed into bunsetsu segments.

3.3 Long Unit Word

The Long Unit Word (LUW) is a word unit based on a bunsetsu. Identification of LUW involves identifying bunsetsu and then dividing each bunsetsu into independent and attached LUWs. For example, in Figure 1, bunsetsu "項目に関する" is divided into an independent LUW "項目 (items)" and attached LUW "に関する (regarding)," which is categorized as adposition even if it contains SUW verb "関する."

4 Corpus

We used the Corpus of Historical Japanese (CHJ; NINJAL 2024), which collects documents from the Nara (AD 710-794) to the Meiji (AD 1868-1912). Bunsetsu and LUW annotations were performed on sampled sentences sampled from the CHJ.

We also used UD-Japanese-GSD[6], a contemporary Japanese corpus, for the model comparison and searching for the best model, because there is a deep-learning-based parser that can output bunsetsu and LUW labels (Matsuda et al., 2022).

Table 1 shows the statistics of both the CHJ and UD-Japanese-GSD. There is not a large difference in the number of sentences in each historical period, while that of UD-Japanese-GSD is one-tenth of them. From the Heian to the Muromachi periods, the number of characters, SUWs, LUWs, and bunsetsu per sentence gradually decreases. In UD-Japanese-GSD, the average numbers of characters and SUWs per sentence are almost the same as those of the Heian period, although the average numbers of LUWs and bunsetsu are less than those of the Heian period.

5 Method

5.1 Bunsetsu and LUW Analyzer Model

Figure 1 shows the architecture of our model. We used joint BI (beginning and inside) tagging-based sequential modeling with a Transformer-based language model (TLM). We combined the

[6] https://github.com/UniversalDependencies/UD_Japanese-GSD

sequential labels of LUW BI, LUW POS, and Bunsetsu BI. For example, the target label of the adposition "は" in Figure 1 is "I-B-ADP," where the first "I" represents the target SUW located intermediate of the bunsetsu, and the second "B-ADP" represents the beginning of the LUW and its POS tag. The total number of target labels is 237 for CHJ and 224 for UD-Japanese-GSD.

We first tokenized each SUW token into subwords instead of tokenizing a sentence directly, to avoid breaking the SUW boundary. We then fed each subword token to the TLM. We added a pooling layer to combine each subword representation produced by the TLM into SUW-level representation. We then fed the pooled SUW-level representations into an additional fully connected layer to output the likelihood of the labels with a softmax activation function. The variants of the pooling layers are as follows:

sum Suppose the j-th subword representation $\boldsymbol{v}_{i,j}$ corresponds to the i-th SUW token output from TLM, the sum pooling $\boldsymbol{u}_i$ is calculated as $\boldsymbol{u}_i = \sum_j \boldsymbol{v}_{i,j}$.

max The max pooling layer takes the max function instead of the summation of the sum pooling.

head The head pooling layer outputs the first subword representation ($\boldsymbol{v}_{i,1}$).

We incorporate SUW POS information into the model in a two-pronged way:

Embedding We concatenated POS embedding with the pooled output $\boldsymbol{u}_i$. The POS embedding was determined through the training.

Incontext We appended a text representing the POS information to each word before subword tokenization. For example, when the SUW "項目 (item)" is tokenized into subwords, the input SUW text representation is "項目 NOUN"[7]. This method increases the number of subword tokens fed into the TLM.

5.2 Evaluation Method

Because our model requires SUW tokens as the input, we feed gold SUWs to the model, throughout the entire evaluation process.

We used span-based precision, recall, and f1 values to evaluate the segmentation of both bunsetsu and LUW. We also used labeled span-based precision, recall, and f1 values for the LUW evaluation. The labeled span-based evaluation is based on a triple (b, e, l) reconstruction score, where b, e, and l represent the start, the end, and the POS labels of the span, respectively.

To evaluate UD-Japanese-GSD, we used the original train, dev, and test sets as intended. We also compare the precision, recall, and f1 values of LUW with the existing parse. Because the prior work tokenized the SUW tokens by a morphological analyzer, we also used predicted SUW tokens by MeCab, instead of the gold SUW tokens.

To evaluate the CHJ samples, we calculated these metrics through five times cross-validations and averaged them to obtain the final scores. We randomly sampled 5% of the sentences from the corpus to create the dev and test sets for each CV. In this procedure, we selected each test set not to overlap.

[7]Though example POS tag is written in English, we add POS tag name in Japanese with sub-tags; "名詞-普通名詞-一般".

	Pooling	P	R	F1
Emb.	sum	.98425	.98264	.98344
	max	.98446	.98446	.98446
	head	**.98532**	**.98456**	**.98494**
Incontext	sum	.98433	.96394	.97403

(a) LUW, span-based

	Pooling	P	R	F1
Emb.	sum	**.97487**	**.97330**	**.97408**
	max	.97228	.97228	.97228
	head	.97348	.97279	.97313
Incontext	sum	**.97478**	.95377	.96416

(b) LUW, labeled span-based

	Pooling	P	R	F1
Emb.	sum	**.97524**	.97459	.97492
	max	.97158	.97350	.97254
	head	.97505	**.97591**	**.97548**
Incontext	sum	.97408	.95488	.96434

(c) Bunsetsu

Table 3: Precision, recall and f1 values of LUW and Bunsetsu tokenization on UD-Japanese-GSD.

	P	R	F1
MeCab + Emb. + sum	0.978	0.978	0.978
Matsuda et al. 2022			
Comainu	0.976	0.969	0.973
SudachiPy + spaCy	0.987	0.985	0.986

Table 4: Span-based LUW score comparison with the previous study.

	Heian			Kamakura			Muromachi		
	P	R	F1	P	R	F1	P	R	F1
Trained on Heian									
LUW span	.99647	.99622	.99635	.98184	.97890	.98036	.90478	.91416	.90945
LUW labeled	.99304	.99279	.99291	.95451	.95165	.95308	.76438	.77231	.76832
Bunsetsu	.96445	.97612	.97025	.93377	.94094	.93734	.74055	.80871	.77313
Trained on Kamakura									
LUW span	.99060	.99147	.99103	.99492	.99452	.99472	.91162	.92650	.91900
LUW labeled	.98252	.98338	.98295	.99089	.99049	.99069	.82257	.83600	.82923
Bunsetsu	.94324	.96250	.95278	.97385	.97997	.97690	.79196	.85138	.82059
Trained on Muromachi									
LUW span	.94672	.95750	.95208	.96079	.95897	.95988	.98913	.98996	.98954
LUW labeled	.88427	.89435	.88928	.91468	.91295	.91381	.98039	.98122	.98080
Bunsetsu	.80727	.86853	.83678	.87293	.89999	.88625	.97810	.97927	.97869

Table 5: Span-based precision, recall, and f1 values on CHJ.

5.3 Hyperparameters

Table 2 lists the hyperparameters. We did not perform an intense hyperparameter search, thus there is a possibility for further performance improvements. Since the number of sentences in CHJ corpora is more than ten times compared to that of in UD-Japanese-GSD, we decreased the total number of epochs and the learning rate when we trained on the CHJ. We used "cl-tohoku/bert-base-japanese-whole-word-masking"[8] for the TLM.

6 Results and Discussions

6.1 Contemporary Japanese

We first compared the model variants using UD-Japanese-GSD, as shown in Table 3. The variant with the **Embedding** and **sum** pooling layers generally performed well. The **head** pooling layer performed well for boundary predictions. This suggests that **sum** pooling provides a better representation of the entire SUW content, while **head** pooling adequately preserves the boundary information.

The variant with **incontext** and the sum pooling achieved the highest precision, but a lower recall value. This is because the **incontext** method increases the number of subword tokens and often exceeds the maximum subword token limit (512) to represent an entire sentence. Table 4 presents a span-based LUW score comparison with that in a previous study (Matsuda et al., 2022). Our model and that of Comainu used MeCab(Kudo et al., 2004) for the SUW tokenization using a UniDic dictionary. The spaCy model uses SudachiPy [9] for SUW tokenizer instead of MeCab. Our model showed an improvement compared to Comainu, while spaCy outperformed the other models. This is because of the difference in the SUW tokenizers.

Because SudachiPy only supports contemporary Japanese, we are supposed to use MeCab for the SUW tokenizer and decided to use **Embedding + sum** pooling model for historical Japanese models.

	Heian	Kamakura	Muromachi
LUW span	.74684	.78141	.77547
labeled	.62969	.68091	.66623
Bunsetsu	.62397	.67525	.67230

(a) F1 values of UD-Japanese-Models on samples of each period.

	Heian	Kamakura	Muromachi
LUW span	.84759	.85769	.88904
labeled	.52768	.56726	.57092
Bunsetsu	.57181	.65828	.75237

(b) F1 values of the models of each period on UD-Japanese-GSD.

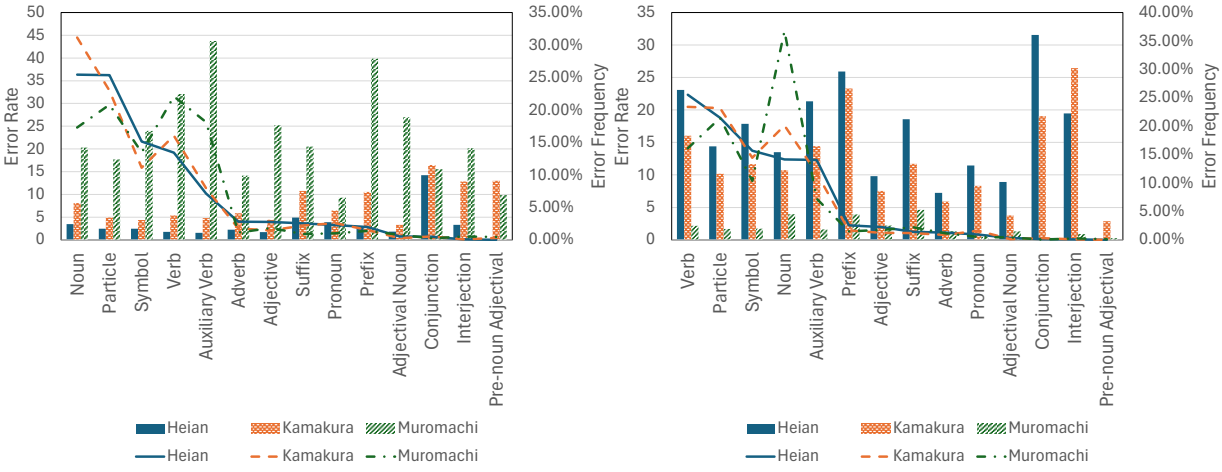
Table 6: Evaluations of zero-shot transfer between contemporary and historical Japanese.

6.2 Historical Japanese

Table 5 lists the overall results for the CHJ. The results evaluated on samples from the same period as

[8] https://huggingface.co/tohoku-nlp/bert-base-japanese-whole-word-masking

[9] https://github.com/WorksApplications/SudachiPy



(a) Trained on the Heian period

(b) Trained on the Muromachi period

Figure 2: POS tags contained in bunsetsu versus error rate and normalized frequency.

during training ranged from 0.976 to 0.996. Thus, our historical models have comparable or even superior results to those of contemporary Japanese (UD-Japanese-GSD), as shown in Table 3. This was because the data size of the CHJ was significantly larger than that of UD-Japanese-GSD. The LUW performances degrades with time, while the bunsetsu segmentation performances increase. As time progresses and vocabulary becomes more complex, it is suggested that styles that are more conscious of syntactic structures such as bunsetsu, increase.

Focusing on the transferability between the CHJ corpora, the model trained on samples of newer periods and applied to older periods yielded higher performance than the reverse case. This is because the vocabulary coverage of the newer samples is larger than that of the older samples. The Heian and Kamakura models work well on samples from each other, however, they do not perform well on samples from the Muromachi period, particularly for labeled LUW and bunsetsu evaluations. This implies drastic grammatical changes occured in Japanese during the Muromachi period.

6.3 Transferability between Contemporary and Historical Japanese

Table 6 shows the transferability performances of contemporary and historical Japanese. In this evaluation, the POS embeddings may not work, because there is a large difference in fine-grained POS categories between contemporary and historical Japanese. Thus, we used the highest level of POS tags for the labeled LUW evaluations. The model trained on UD-Japanese-GSD performed similarly in each period (Table 6a). However, the performances of the models on samples from each period increased with time, specifically for the bunsetsu segmentation. This indicates that the syntactic structure of sentences gradually approaches modern syntactic structures over time, while the morphology of LUW is not as high.

6.4 Grammatical Changes during the Muromachi Period

Figure 2 plots the error rates of bunsetsu containing the SUW of a particular POS tag. Figure 2 presents the results of the models trained on samples from the Heian and Muromachi periods. We also plotted the normalized error frequency corresponding to each POS tag for all errors in the same period in Figure 2.

The model trained on Heian period data exhibited a particularly higher error rate when it predicted bunsetsu containing auxiliary verbs or verbs when evaluated on samples from the Muromachi period. This tendency was also observed when samples from the Heian period were evaluated using the Muromachi model. This indicates that there may have been significant changes in sentence endings that usually contained both verbs and auxiliary verbs.

When evaluating samples from the Muromachi period using the Heian model, the error frequency relatively increased in bunsetsu-containing nouns compared with the reverse scenario. This is because the newer model partially contains old vocabularies.

In both cases, the bunsetsu-containing particles resulted in a high error rates and frequencies.

Gold (Heian) and the Kamakura model prediction:													
En	One doesn't do such things, there will surely be regrets												
Ja	さる	わざ	せ	ず	は	、	恨むる	こと	も	あり	な	む	など
	such	things	do	not			regrets			there be	surely	will	
LUW	V	N	V	A	P	S	V	N	P	V	A	A	P
The Muromachi model prediction:													
LUW	C	V		A	P	S	V	N	P	V	V	A	P

Table 7: An example of bunsetsu and LUW analysis. V, N, A, P, S, and C stand for verb, noun, auxiliary verb, particle, symbol, and conjunction, respectively. Vertical bars represent bunsetsu boundaries.

Evaluated on	Heian		Muromachi	
	C	R	C	R
Sentence-ending	95	13.67	1835	43.74
Adverbial	1521	15.41	462	27.26
Case-marking	11577	10.46	7395	11.22
Binding	5216	11.55	2026	10.27
Conjunctive	2966	15.99	1615	12.49

Table 8: Error counts (C) and error rate (R) of bunsetsu ending with a particle. We show a result of the Muromachi model evaluated on data in the Heian period, and vice versa.

Case Study: Verbs and Auxiliary Verbs Table 7 presents a sample sentence from the Heian period data and the outputs of our models. The Japanese space-separated tokens in Table 7 are SUW tokens. In this case, the LUWs and SUWs are identical. V, N, A, P, S, and C denote verb, noun, auxiliary verb, particle, symbol, and conjunction, respectively. Vertical bars represent the bunsetsu boundaries.

The Heian and the Kamakura models output perfect LUW and bunsetsu boundaries, respectively. The first word "さる (saru; do such)" is a verb, however, it is often used as an adversative conjunction, and thus the Muromachi model misclassified it as a conjunction. The second verb "せ (se; do)" often composes a LUW with an antecedent noun. The first noun "わざ (waza; thing)" has several senses, such as "ceremony" and "technique"; thus "わざせ" is misunderstood as "doing a ceremony" or "doing the technique" by the Muromachi model. This is because a case marker "を (wo; objective)" is required just after "わざ" to retain the meaning in the Muromachi period.

Both "さる" and "せ" are common words; thus, it is conceivable that the grammaticalization of those words was progressing during the Muromachi period. Since the verbs and auxiliary verbs are often contained in mispredicted bunsetsu in Figure 2, the grammaticalization of those words would be a major part of the grammatical changes.

The auxiliary verb "な (na; complete)" is misclassified as a verb. This may be because the expression "なむ" became less common in the Muromachi period.

Analysis of Particles Table 8 lists the error counts and error rates of bunsetsu prediction when the target bunsetsu ends with a particle for all particle subcategories. During the Heian period, adverbial particles were frequently used. However, during the Muromachi period, they became less common. Conversely, while there were a few examples of sentence-ending particles in the Heian period, they became commonly used in the Muromachi period [10]. The error rates of bunsetsu prediction ending with these particles significantly increased when the Heian model was applied to data from the Muromachi period. This could be because new usages for these particles emerged during the Muromachi period alongside the changes in verb conjugation forms, which often appear with the sentence-ending particles.

7 Conclusion

This study focuses on providing a deep learning-based (or LLM-based) bunsetsu, which is a minimal phrase in Japanese, and Long Unit Words parser for the Heian period (AD 794-1185) to the Muromachi period (AD 1336-1573) and evaluating its performances.

We model the parser as a joint sequential label that combines the bunsetsu BI tag, LUW BI tag, and LUW POS tags for each SUW token. We used the transformer-based language model to output an SUW token representation by taking the appropri-

[10]The samples of the Muromachi period are mainly informal conversations, which used sentence-ending particles frequently. https://clrd.ninjal.ac.jp/chj/muromachi-en.html

ately pooling of the subword representations for the last layer of the transformer. We trained our models on the corpora of each period, including contemporary and historical Japanese.

The results ranged from 0.976 to 0.996 in the f1 value for both bunsetsu and LUW reconstructions indicating that our models achieved comparable performance to models trained on a contemporary Japanese corpus.

Through the statistical analysis and case studies comparing each period, the bunsetsu estimation can be influenced by the grammaticalization of morphemes.

In the future, we will expand the applicable periods. We will build a syntactic parser by annotating the dependencies between bunsetsu segments.

Acknowledgments

This research was supported by JSPS KAKENHI grant JP22K12145, as well as the National Institute for Japanese Language and Linguistics Joint Research Projects "Empirical Computational Psycholinguistics Using Annotation Data." and "Extending the Diachronic Corpus through an Open Co-construction Environment."

References

Masayuki Asahara, Hiroshi Kanayama, Takaaki Tanaka, Yusuke Miyao, Sumire Uematsu, Shinsuke Mori, Yuji Matsumoto, Mai Omura, and Yugo Murawaki. 2018. Universal Dependencies version 2 for Japanese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yasuharu Den, Junpei Nakamura, Toshinobu Ogiso, and Hideki Ogura. 2008. A proper approach to Japanese morphological analysis: Dictionary, model, and evaluation. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Shunsuke Kozawa, Kiyotaka Uchimoto, and Yasuharu Den. 2014. Bccwj ni motozuku cho-tani kaiseki tsuru comainu (in japanese)(long unit word analysis tool comainu based on bccwj). *The Annual Meeting of the Association for Natural Language Processing*.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237, Barcelona, Spain. Association for Computational Linguistics.

Hiroshi Matsuda. 2020. Ginza - practical japanese nlp based on universal dependencies. *Journal of Natural Language Processing*, 27(3):695–701.

Hiroshi Matsuda, Mai Omura, and Masayuki Asahara. 2022. Ud japanese ni motoduku kokugo-ken chotani kaiseki-kei no kochiku(in japanese) construction of a long unit word analysis system for japanese based on ud japanese. *The Annual Meeting of the Association for Natural Language Processing*.

NINJAL. 2024. Corpus of historical japanese.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Mai Omura and Masayuki Asahara. 2018. UD–Japanese BCCWJ: Universal Dependencies annotation for the Balanced Corpus of Contemporary Written Japanese. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 117–125, Brussels, Belgium. Association for Computational Linguistics.

Mai Omura, Aya Wakasa, and Masayuki Asahara. 2023. Universal dependencies for japanese based on long-unit words by ninjal. *Journal of Natural Language Processing*, 30(1):4–29.