UniDic for Early Middle Japanese: a Dictionary for Morphological Analysis of Classical Japanese

Toshinobu Ogiso*†, Mamoru Komachi†, Yasuharu Den‡, Yuji Matsumoto†
*Department of Corpus Studies, National Institute for Japanese Language and Linguistics (NINJAL)
†Graduate School of Information Science, Nara Institute of Science and Technology (NAIST)
‡Faculty of Letters, Chiba University

10-2, Midori-cho, Tachikawa-shi, Tokyo JAPAN 190-8561
E-mail: togiso@ninjal.ac.jp, komachi@is.naist.jp, den@cogsci.l.chiba-u.ac.jp, matsu@is.naist.jp

Abstract

In order to construct an annotated diachronic corpus of Japanese, we propose to create a new dictionary for morphological analysis of Early Middle Japanese (Classical Japanese) based on UniDic, a dictionary for Contemporary Japanese. Differences between the Early Middle Japanese and Contemporary Japanese, which prevent a naïve adaptation of UniDic to Early Middle Japanese, are found at the levels of lexicon, morphology, grammar, orthography and pronunciation. In order to overcome these problems, we extended dictionary entries and created a training corpus of Early Middle Japanese to adapt UniDic for Contemporary Japanese to Early Middle Japanese. Experimental results show that the proposed UniDic-EMJ, a new dictionary for Early Middle Japanese, achieves as high accuracy (97%) as needed for the linguistic research on lexicon and grammar in Japanese classical text analysis.

Keywords: Morphological Analysis, Classical Japanese, Early Middle Japanese, Historical Corpus of Japanese

1. Background

Recently, the use of corpus linguistics has become popular among Japanese linguists. To facilitate further research on corpus linguistics, the National Institute for Japanese Language and Linguistics (NINJAL) has compiled one of the largest Japanese corpora, the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa et al., 2010). Following the same line of research, a diachronic corpus of Japanese is currently under construction.

Since corpus linguistics heavily relies on word-segmented corpora, it is important to have morphological annotations for the corpus that is the object of study. However, morphological annotations do not come for free, and thus an automatic morphological analyzer is desired for Japanese corpus linguists. To implement highly accurate and effective morphological analyzers, a carefully constructed wide-coverage dictionary is necessary. It is essential for statistical and machine learning-based approaches to be successful. For example, the state-of-the-art Japanese morphological analyzer MeCab (Kudo et al., 2004) is trained with an electronic dictionary called UniDic [1] on a manually annotated BCCWJ. In UniDic, all entries are based on the definition of short unit word (SUW), which provides word segmentation in uniform size suited for linguistic research. UniDic also achieves high performance in many text genres including literature, spoken texts, and so on (Den et al., 2007).

However, the original UniDic is only for the Contemporary Japanese (CJ). We conducted preliminary experiments of morphological analysis of literature written in Early Middle Japanese (EMJ) by adopting the state-of-the-art morphological analyzer MeCab with contemporary dictionaries. It turned out that its accuracy on EMJ was considerably lower than the reported accuracy for newswire texts, and completely inadequate for Japanese linguists. One of the reasons is that because there was a massive change in writing style in the Meiji era (1868-1912).

Early Middle Japanese is a historical stage of the Japanese language used in the Heian period (A.D. 794 - 1185). In the Heian period, various styles of Japanese literature such as *monogatari* (tales) and *nikki bungaku* (diary literature) appeared for the first time in history. *Waka* (native Japanese poetry) also flourished at this time. For example, masterpieces such as the *Tale of Genji*, the *Tosa Diary*, and the *Kokin Waka-shū* poetry anthology were written in this era, to name a few. Therefore, a morphological analysis of EMJ is especially useful for Japanese historical linguists.

As the first step toward rich annotation of linguistic information for historic texts in the diachronic corpus, we propose to start with building an electronic dictionary for morphological analysis adapted for EMJ. Morphological analysis is one of the fundamental annotations for construction of a full-scale corpus.

The rest of this paper is organized as follows. Section 2 describes characteristics of Early Middle Japanese. Section 3 explains how we built the UniDic for Early Middle Japanese. Section 4 compares the UniDic for Early Middle Japanese with other dictionaries to show its effectiveness. Section 5 presents conclusions and suggests future direction.

2. Linguistic Characteristics of Early Middle Japanese

Early Middle Japanese has various characteristics that distinguish it from CJ in several linguistic fields: lexicon, morphology, syntax, orthography and pronunciation. We

[1] http://download.unidic.org/

will briefly describe the differences between CJ and EMJ in terms of corpus linguistics.

2.1 Lexical Differences

The Japanese lexicon mainly consists of three types of words: *wago*, *kango* and *gairaigo*. *Wago* are words of Japanese origin which had existed before *kango* were introduced from China. *Kango* are words of Chinese origin which were imported from China or created in Japan using *kanji* (Chinese characters). *Gairaigo* are foreign words not originating from Chinese, usually transliterated and written in *Katakana*. These word types are called *"goshu"*. In CJ, approximately 18% to 70% of words used in texts are *kango* (in SUW). On the contrary, in the literary text in EMJ only 1% to 5% of words are *kango*. This fact suggests that numerous *kango* words have been newly imported or created and many *wago* words have become obsolete, even though most of the basic words in EMJ are *wago* and still remain the same today. Thus, dictionaries for CJ tend to lack outdated but essential words.

2.2 Morphological Differences

Conjugation type has changed throughout the history of Japanese language. For example, conjugation of verb "*kuru* 来る" (come) and adjective "*akai* 赤い" (red) have changed as below (Table 1).

	Conjugation	EMJ	CJ
kuru 来る (*v.*come)	*mizen* (irrealis)	ko	ko
	ren'yō (continuative)	ki	ki
	shūshi (terminal)	ku	kuru
	rentai (attributive)	kuru	kuru
	izen (realis) */ katei* (hypothetical)	kure	kure
	meirei (imperative)	ko	koi
akai 赤い (*adj.*red)	*mizen* (irrealis)	akaku (akakara)	akakaro
	ren'yō (continuative)	akaku (akakari)	akaku akakat-
	shūshi (terminal)	akasi	akai
	rentai (attributive)	akaki (akakaru)	akai
	izen (realis) */ katei* (hypothetical)	akakere	akakere
	meirei (imperative)	(akakare)	akakare

Table 1. Differences of Conjugation

Though most lexical entries of verbs had already been included in the UniDic dictionary and most of the conjugations in EMJ can be formed by derivation, the conjugation table had to be modified for EMJ. Because there are many irregularly changed words and many contemporary words not used in EMJ, we had to check all derived entries.

Moreover, this difference in conjugation type affects word bigram probability, since conjugations of verbs are abundant in texts. Thus, a naïve application to EMJ is not practical for the part-of-speech tagging model learned from CJ.

2.3 Grammatical Differences

Although the word order of EMJ is almost the same as that of CJ, function words such as particles and auxiliary verbs have changed considerably over time. For example, the most frequently used auxiliary verbs in EMJ, such as *"mu"*, *"beshi"*, *"keri"*, are no longer used today. For this reason, corpora of CJ are not appropriate for machine learning-based approaches to morphological analysis of EMJ.

2.4 Orthographic Differences

There are many orthographic differences between EMJ and CJ texts. Usages of *kana* and *kanji* characters are the most significant differences. Table 2 shows the examples of these differences.

	Word (meaning)	CJ	EMJ
Kana Usage	*koe* (*n.* voice)	こえ	こゑ
	omou (*v.* think)	おもう	おもふ
Kanji Usage	*kuni (n. country)*	国	國
	kuru (v. come)	来る	來る
Kana and Kanji	*au (v. meet)*	会う	會ふ

Table 2. Differences of *kana* and *kanji* Orthography

In EMJ, words written in the *kana* orthography were spelled in *Rekishi Kanazukai* (historical kana usage) based on the pronunciations at the time. *Rekishi Kanazukai* was the mainstream orthography until the *Gendai Kanazukai* (modern *kana* usage) was introduced in 1946. Because most morphological analyzers do not canonicalize these usages, they fail to analyze these characters correctly.

Furthermore, there are some old *kanji* characters not present in CJ. Since EMJ contains old variants of *kanji*, these characters deteriorate the performance of morphological analysis if the dictionary only includes the newer counterparts.

There is a further complication: Old *kanji* and different *kana* usage are often used compositely.

3. Making the UniDic for Early Middle Japanese

In order to overcome the problems stemming from the differences between the Contemporary Japanese and the Early Middle Japanese mentioned above, we decided to build a new dictionary and a corpus especially for EMJ. We used two approaches: One is to expand entries of the contemporary UniDic dictionary, and the other is to annotate a new corpus of EMJ as training data for morphological analysis.

3.1 Extension of Dictionary Entries

Starting from the existing UniDic, we extended word entries to cope with the problem of lexical, morphological and orthographic differences.

As was mentioned above, UniDic is an electronic dictionary designed for linguistic use. UniDic is structured with layered entries to treat words flexibly depending on the purposes of researchers.

Figure 1 exemplifies the structured word indexes of UniDic. The Lemma layer is prepared to treat words at abstract lemmatized level, like the entries of the general dictionary. The Form layer is prepared to distinguish allomorphs and different conjugations. Specification of conjugations type is held in this layer. The Orthographic layer is prepared to distinguish orthographic variants.

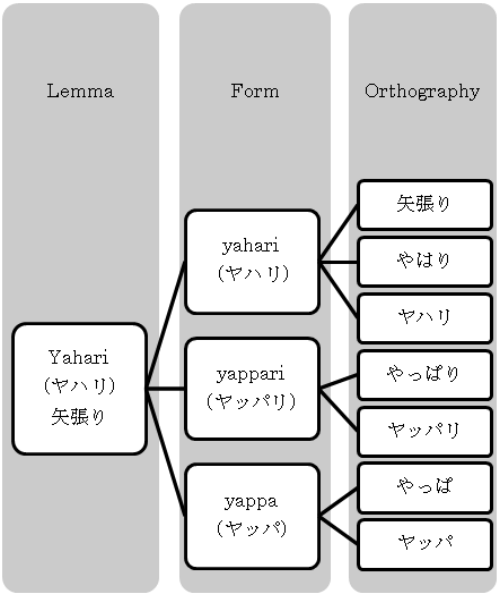


Figure 1. Hierarchical Structure of UniDic

This structure helped us to add new entries in each level. For example, morphological differences like word forms or conjugations are handled in the Form level, and orthographic differences such as *kana* usage and *kanji* variants are dealt with in the Orthography level.

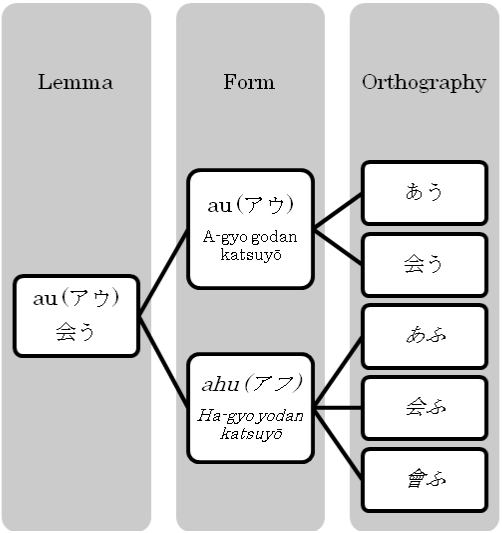


Figure 2. Extensions of EMJ Word Entries

Figure 2 shows the extensions of word entries for EMJ. In this figure, the Form "*ahu*" is added to annotate old conjugation forms in EMJ corresponding to the CJ word "*au* 会う" (meet) . Likewise, old orthographic forms of "*ahu*" such as "あふ" and "會ふ" are added under the form.

Each conjugation form is generated automatically by applying the inflection table prepared for EMJ.

We added approximately 20,000 entries to cope with the lexical, morphological, and orthographic differences. Rules of newly added entries for EMJ are summarized in Ogura et al., (2012).

3.2 Training Corpus of Early Modern Japanese

To remedy the issues of morphological and syntactic differences, we manually annotated a corpus of EMJ containing 271,000 words (SUWs) to produce training and test corpora. Table 3 summarizes the texts we selected. This corpus contains major styles of Japanese literature such as *monogartari* and *nikki bungaku*, and thus serves as the fundamental resource for EMJ.

Text	Number of Words
(A part of) The Tale of Genji *(Genji Monogatari)*	172,929
The Diary of Lady Murasaki *(Murasaki Shikibu Nikki)*	20,350
The Tosa Diary *(Tosa Nikki)*	7,948
As I Crossed a Bridge of Dreams *(Sarashina Nikki)*	16,656
The Tales of Ise *(Ise Monogatari)*	14,624
The Tales of Yamato *(Yamato Monogatari)*	26,478
The Tale of the Bamboo Cutter *(Taketori Monogatari)*	12,136

Table 3. Annotated Corpus of EMJ

3.3 Configuration of Analyzer

MeCab is a morphological analyzer based on CRF (Lafferty et al., 2001) and achieves state-of-the-art performance in Contemporary Japanese morphological analysis. One of the main advantages of the tool is that its feature template is flexibly designable. We added the feature of archaic particles, affixes, and auxiliary verbs in order to address the problem of grammatical differences between EMJ and CJ. Furthermore, the *goshu* features are also added to correspond with the lexical differences. *Goshu* features have been used for the original UniDic (for CJ) and it is confirmed that they are effective (Den et al., 2007). MeCab can automatically learn feature weights for UniDic from an annotated corpus of EMJ to build a morphological analyzer.

As local context, MeCab uses part-of-speech-level bigram for general words to avoid sparseness, with the only exception of function words such as particles or affixes,

which use word-level bigram (lexicalization). In the setting of UniDic-EMJ, word-level bigram is used for archaic particles, auxiliary verbs and affixes, in place of function words of CJ. All the other configurations of the analyzer basically remain at the same setting as is used for CJ.

4. Evaluation of the UniDic for Early Middle Japanese

4.1 Experimental Settings

We evaluated the performance of the UniDic for EMJ version 0.6. The test data contains 27,100 words (SUWs) of randomly sampled sentences (10% of the annotated corpus). Note that although the test data was not used as training corpus, it contained no words unknown by the dictionary.

The evaluations were carried out in four levels. Level 1 is the accuracy of word segmentation. Level 2 is the accuracy of part-of-speech tagging for items correct at Level 1. Level 3 is the accuracy of lemmatization for items correct at Levels 1 and 2. Level 4 is the accuracy of distinction of allomorphs for items correct at all other levels. Table 4 shows the number of correct words in the analyzed texts and corresponding degrees of performance. for the four levels

The accuracy of Level 3, which is mainly used by linguists, isz approximately 97%. This number is not so much inferior in comparison with the accuracy of the morphological analysis dictionary of CJ (approximately 98%). Although it depends on the purposes of the research in question, 97% accuracy is sufficient for a variety of historical linguistic studies.

4.2 Comparison with Other UniDics

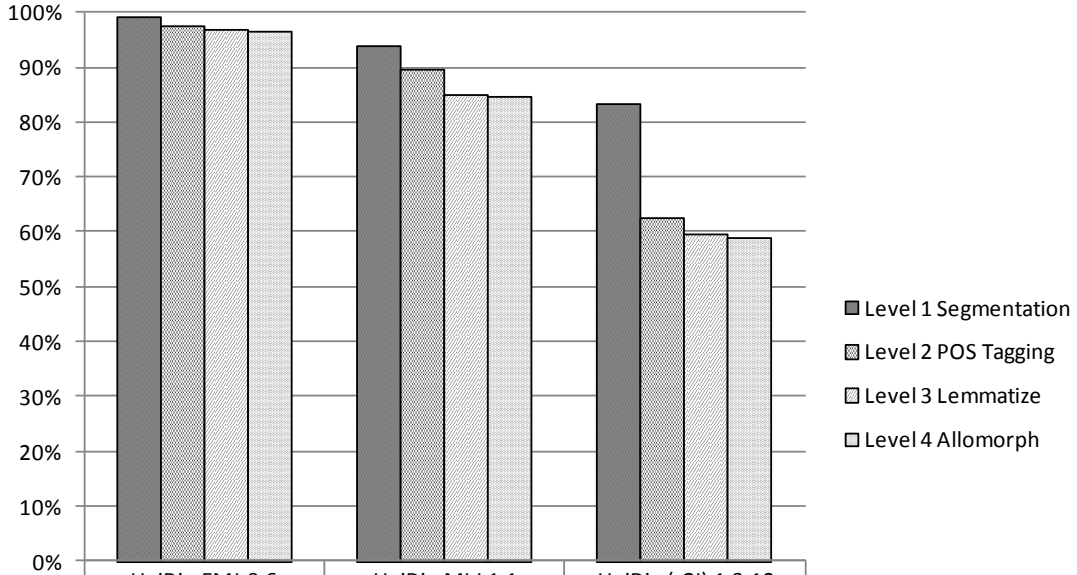
We compared the performance of UniDic-EMJ with the original UniDic and UniDic-MLJ (*Kindai-Bungo* UniDic) in the analysis of Japanese classical texts. Original UniDic (UniDic-CJ) does not contain obsolete words. UniDic-MLJ is a morphological dictionary for Modern Literary Japanese (literary style texts in Meiji Era). Although UniDic-MLJ contains almost the same lexicon as UniDic-EMJ, it is trained on a different corpus.

	Level 1	Level 2	Level 3	Level 4
Input words	25,535			
Output words	25,524			
Correct words	25,361	24,939	24,759	24,649
Recall	0.99319	0.97666	0.96961	0.9653
Precision	0.99361	0.97708	0.97003	0.96572
F-value	0.99340	0.97687	0.96982	0.96551

Table 4. Numbers of Correct Words and Accuracy

The test data for this comparison is the same as the data used in Table 3. This test corpus is outside of the data domain of both UniDic-CJ and UniDic-MLJ, and thus it is no wonder they do not perform well on this data set. However, these two had been the only available dictionaries for EMJ until UniDic-EMJ was built.

Figure 3 shows the performance of the three variants of UniDics using the same criteria as Table 3. As you can see,



	UniDic-EMJ 0.6	UniDic-MLJ 1.1	UniDic (-CJ) 1.3.12
Level 1 Segmentation	0.99340	0.94003	0.83265
Level 2 POS Tagging	0.97687	0.89464	0.62586
Level 3 Lemmatize	0.96982	0.85030	0.59499
Level 4 Allomorph	0.96551	0.84554	0.59002

Figure 3. Performance Comparison with Other UniDics

UniDic-EMJ achieved the best performance. UniDic-EMJ outperformed UniDic-CJ for POS Tagging, Lemmatization and Allomorph by a large margin. This clearly demonstrates the effectiveness of building a tailored dictionary for a specific period for historical text at hand.

4.3 Error Analysis

We carried out an error analysis of the morphological analysis using UniDic-EMJ. At Level 1, complex compound words are divided into a set of two or more simple words. For example, "*tabikasanaru*" (repeat) is divided into "*tabi*" (time) and "*kasanaru*" (overlap), and "*kataharaitasi*" (disgusting) is divided to "*katahara*" (side) and "*itasi*" (painful). At Level 2, there are many mistakes in distinguishing short function words of the same form. One of the most frequent words, "*ni*", can be one of three different parts of speech: dative case marker, conjunction particle or a conjugated form of the copula "*nari*". Errors also occur in the distinction between a noun derived from a verb and the original verb: for example, the noun "*wakare*" (parting, separation) and the *ren'yō* (continuative) form of the verb "*wakareru*" (part, separate). Some verbs realize two different conjugational forms with the same surface form and ambiguities such as these also caused a large number of errors. For example, both the *shūshi* (terminal) conjugation and the *rentai* (attributive) conjugation of the verb "*tatu*" (stand) take the form "*tatu*" and are written in identical ways.

At Level 3, there are many errors in identifiying *wago* words of kindred meaning expressed by the same *kanji*. For example, "*ne*" and "*oto*" (sound) are both written "音"; "*sita*" and "*simo*" (under) are both written "下"; "*toko*" and "*yuka*" (bed or floor) are both written "床", and so on. Some errors were failures to recognize the distinction between *wago* and *kango* written in the same *kanji*, such as *wago* "*ama*" and *kango* "*ni*" (written "尼"). However, as the result of using *goshu* features, such errors were reduced. At Level 4, a large number of errors were due to variations of forms produced by *rendaku* (voicing of the initial consonant).

All these errors are hard to distinguish even for humans. Automatic morphological analysis using UniDic-EMJ has already accomplished a level of accuracy as high as that of ordinary non-experts.

5. Conclusions and Future Work

We have constructed an electronic dictionary for morphological analysis of Early Middle Japanese (Classical Japanese), which can analyze Japanese classical texts with high accuracy. Its accuracy (97%) is considered to be high enough for linguistic research on lexicon and grammar. UniDic-EMJ is now freely available at our webpage [2]. Several reports on the development of UniDic-EMJ, software tools in association with UniDic-EMJ, and linguistic studies using UniDic-EMJ are summarized in Ogiso et al., (2012).

For the compilation of a Japanese diachronic corpus, we must prepare more dictionaries for other types of Japanese language: other times and genres. One highly needed resource is a dictionary for colloquial Early Modern Japanese. We are planning to build a new UniDic to analyze texts of this type.

6. Acknowledgements

This work is partially supported by the collaborative research project "Study of the history of the Japanese language using statistics and machine-learning" carried out at the National Institute for Japanese Language and Linguistics.

7. References

Yasuharu Den, Junichi Nakamura, Toshinobu Ogiso, and Hideki Ogura. (2008). A proper approach to Japanese morphological analysis: Dictionary, model, and evaluation. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008)*, Marrakech, Morocco, pp. 1019-1024.

Yasuharu Den, Toshinobu Ogiso, Hideki Ogura, Atsushi Yamada, Nobuaki Minematsu, Kiyotaka Uchimoto and Hanae Koiso. (2007). The development of an electronic dictionary for morphological analysis and its application to Japanese corpus linguistics (in Japanese). *Japanese Linguistics*, 22: pp.101–123.

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*, pp. 230–237, Barcelona, Spain.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pp. 282–289, Williamstown, MA.

Kikuo Maekawa, Makoto Yamazaki, Takehiko Maruyama, Masaya Yamaguchi, Hideki Ogura, Wakako Kashino, Toshinobu Ogiso, Hanae Koiso and Yasuharu Den. (2010). Design, compilation, and preliminary analyses of balanced corpus of contemporary written Japanese. In *Proceedings of the 7th Language Resources and Evaluation Conference (LREC 2010)*. Valletta, Malta, pp. 1483-1486.

Hideki Ogura, Tetsuya Sunaga, Toshinobu Ogiso. (2012). *Rules of Short Unit Words for UniDic-EMJ*. (in Japanese), Research Report of the Grants-in-Aid for Scientific Research (Project Number: 21520492).

Toshinobu Ogiso, Hideki Ogura, Makiro Tanaka, Asuko Kondo, Yasuharu Den. (2012). *Development of an Electronic Dictionary for Morphological Analysis of Classical Japanese*. (in Japanese) Research Report of the Grants-in-Aid for Scientific Research (Project Number: 21520492).

[2] http://www2.ninjal.ac.jp/lrc/index.php?UniDic