Word Delimitation Issues in UD Japanese

Word Delimitation Issues in UD Japanese

Mai Omura
NINJAL, Japan

Aya Wakasa
NINJAL, Japan

Masayuki Asahara
NINJAL, Japan

{mai-om, awakasa, masayu-a}@ninjal.ac.jp

Abstract

This article discusses word delimitation issues in Universal Dependencies (UD) Japanese. The Japanese language is morphologically rich and does not use white space to delimit words. Word delimitation is an important issue in the development of language resources. Even though UD defines the base unit word using *syntactic words*, UD Japanese utilises **Short Unit Words (SUW)**, which are nearly the same as *morphemes*, the base unit word. We developed another word delimitation version of UD Japanese resources that uses **Long Unit Words (LUW)** as the base unit word, which can be regarded as *syntactic words* in Japanese. We then evaluated their reproducibility through publicly available language resources. The results show that the word delimitation and dependency structure of LUW-based UD Japanese reproduce the results using SUW-based UD Japanese. However, the lemmatisation of LUW is still more complex than that of SUW for a morphologically rich language.

1 Introduction

Universal Dependencies (Nivre et al., 2016) (UD) define the base unit word of dependency annotation as *syntactic words*. Languages with white spaces in their word delimitation tend to utilise space as word boundaries. However, languages that do not use white space to delimit words (e.g. Chinese (Xia, 2000; Leung et al., 2016) and Korean (Chun et al., 2018)) present issues in defining their *syntactic words*. For example, while UD Chinese defines word delimitation using the available word-segmented corpus, UD Classical Chinese (Yasuoka, 2019) did not define syntactic words and utilised characters as the word unit. Even when we use characters as the base unit, the lexicon size is approximately 7,000 for simplified Chinese characters and 13,000 for traditional Chinese characters.

Murawaki (2019) pointed out that the preceding versions of the UD Japanese utilise morphemes as the base unit. The word delimitation is based on the **Short Unit Word** (短単位: hereafter **SUW**), defined by the National Institute for Japanese Language and Linguistics, Japan (hereafter NINJAL). Currently, we have the SUW-based word lexicon UniDic (Den et al., 2007) with 879,222 entries and morpheme-based word embeddings NWJC2vec (Asahara, 2018) with 1,589,634 entries for Japanese. The large size of the lexicon is because Japanese is a morphologically rich language. When we use a longer word unit as the base unit, the lexicon size is larger, and the token type ratio becomes larger. Practically, SUW-based UD Japanese resources can be developed with less effort from publicly available language resources. Thus, we had utilised SUW as the base unit word in UD Japanese.

We newly developed another version of UD Japanese with **Long Unit Word** (長単位: hereafter **LUW**) delimitation. The LUW definition by NINJAL can be regarded as *syntactic words* in Japanese. Even though LUW delimitation is appropriate for the base unit words of UD, the cost of LUW-based corpus development is much higher than that of SUW-based corpus development. Furthermore, the reproducibility of LUW-based UD Japanese should be investigated.

This paper presents LUW-based UD Japanese language resources. We also present the reproducibility of LUW-based UD Japanese structures using currently available tools and language resources. The remainder of this paper is organised as follows. Section 2 presents word delimitation in Japanese, including

currently available tools and language resources. Section 3 presents LUW-based UD Japanese language resources. Sections 4 and 5 present an experimental evaluation of their reproducibility. Finally, Section 6 concludes the paper.

2 Word Delimitation in Japanese

2.1 Japanese word delimitation standards by NINJAL

Min. Unit	‖ 全 ‖ 学 ‖ 年 ‖ に ‖ わたっ‖ て ‖ 小 ‖ 学 ‖ 校 ‖ の ‖ 国 ‖ 語 ‖ の ‖
	教 ‖ 科 ‖ 書 ‖ に ‖ 大 ‖ 量 ‖ の ‖ 挿し ‖ 絵 ‖ が ‖ 用い ‖ られ ‖ て ‖ いる ‖
SUW	‖ 全 ‖ 学年 ‖ に ‖ わたっ‖ て ‖ 小 ‖ 学校 ‖ の ‖ 国語 ‖ の ‖
	教科 ‖ 書 ‖ に ‖ 大量 ‖ の ‖ 挿し絵 ‖ が ‖ 用い ‖ られ ‖ て ‖ いる ‖
LUW	‖ 全学年 ‖ にわたって ‖ 小学校 ‖ の ‖ 国語 ‖ の ‖
	教科書 ‖ に ‖ 大量 ‖ の ‖ 挿し絵 ‖ が ‖ 用い ‖ られ ‖ ている ‖
Bunsetsu	‖ 全学年にわたって ‖ 小学校の ‖ 国語の ‖
	教科書に ‖ 大量の ‖ 挿し絵が ‖ 用いられている ‖
(*romanisation*)	‖ zen gakunen ni watatte ‖ syou gakkou no ‖ kokugo no ‖
	kyoukasho ni ‖ tairyou no ‖ sashie ga ‖ mochii rare te iru ‖
(*gloss*)	‖ for all school years ‖ elementary school-GEN ‖ Japanese language-GEN ‖
	textbooks-DAT ‖ many ‖ picture-PL-SBJ ‖ use-PASS-PRET ‖

Translation: *Many pictures are used in elementary school textbooks for all school years.*

Figure 1: Example of Minimum Unit, SUW, LUW, and Bunsetsu in BCCWJ PB33_00032

NINJAL defines several word delimitation standards: Minimum Unit (最小単位), SUW, LUW, and *Bunsetsu* (文節), shown in Figure 1 (Den et al., 2008).

The Minimum Unit standard (最小単位) is defined by word types. Japanese has the following word types: Chinese-origin words (漢語), Japanese-origin words (和語), Loan words other than Chinese-origin words (外来語), Symbols (記号), Numerals (数値表現), and Proper nouns (固有名詞). Chinese-origin words are split into individual characters. Japanese-origin words are split into their shortest units. Loan words other than Chinese-origin words are split into the original shortest unit. Numerals are split into the pronounceable decimal digits. For example, "1076" is split into "千" (*sen*; one thousand), "七十" (*nanajuu*; seventy) and "六" (*roku*; six). Symbols are split into individual characters. Proper nouns are split into their shortest units.

SUWs are defined by the Minimum Unit standards for their word type: Minimal Unit lexicon MORPH $= \{m_1, \ldots\}$ with word types WORDTYPE $= \{wt_{m_1}, \ldots\}$. SUW is defined as follows: If a word is a Minimal Unit (word $\in$ MORPH), then word is SUW. If a word is split into two Minimal Units m_A, m_B and their word types are the same ($wt_{m_A} = wt_{m_B}$), then word is SUW. Note that if a word is split into more than two Minimal Units, the word is not SUW.

Parts of speech (POSs) can be assigned to SUWs. In Japanese, verbs, adjectives, and auxiliary verbs have conjugations. These three POSs have conjugation types (CTYPEs) and conjugation forms (CFORMs). SUW can be categorised as dependent (付属語) and independent words (自立語) by POS. These two correspond to functional and content words in UD. Postposition (助詞), auxiliary verb (助動詞), prefix (接頭辞) and suffix (接尾辞), are categorised as dependent words. The conjugation types are defined by their conjugation patterns, such as class-5 verbs (五段), class-1 verbs (一段), and irregular verbs. The conjugation forms are defined as irrealis form (未然形), conjunctive form (連用形), and so on.

LUW is defined by the Bunsetsu (文節) delimitation. Before defining the LUW delimitation, we define the Bunsetsu delimitation. Bunsetsu is a base phrase in Japanese, which is similar to *eojeol* (語節) in Korean. Bunsetsu composes one compound independent word and dependent words, such as prefix morphemes, postpositions, and auxiliary verbs. Bunsetsu-based Japanese dependency structures have the following properties useful when developing dependency parsers: They are (a) mostly projective,

(b) strictly head-final, and (c) easily produce Bunsetsu delimitation by chunkers. Bunsetsu-based dependency parsers have mainly developed been in the Japanese natural language processing fields (Kudo and Matsumoto, 2002; Kawahara and Kurohashi, 2006). However, since Bunsetsu is a base phrase, POS is not assigned to the Bunsetsu unit.

LUW delimitation is defined as constituents in the Bunsetsu. Because LUWs have their POS and morphological features of conjugation, we can use LUW as the *syntactic words* in the UD standard. One compound-dependent word with prefix morphemes is the semantic head LUW word in Bunsetsu. Most of the SUWs of postpositions, auxiliary verbs, and suffixes are regarded as one LUW. However, NINJAL LUW delimitation defines multi-word functional expression as one LUW.

2.2 Language resource availability

	Lexicon	Word Segmenter	Segmented Corpus	Word Embeddings	TTR
Characters	UTF-8 charset	buildable	buildable	buildable	0.00004
Minimal Unit	N/A	N/A	N/A	N/A	N/A
SUW	UniDic	MeCab	BCCWJ	NWJC2vec	0.00176
LUW	N/A	Comainu	BCCWJ	N/A	0.02922
Bunsetsu	N/A	Comainu	BCCWJ	N/A	0.22221

Table 1: Language resource availability

This section presents availability of the language resources. Table 1 shows the language resource availability for the delimitation. Characters can be produced by simple scripts. Because the Minimal Unit is the unit to determine SUW delimitation manually, there is no publicly available resource for doing so. UniDic [1] is an SUW-based lexicon which can be used in the word segmenter MeCab (Kudo et al., 2004) [2]. Neither LUW nor Bunsetsu lexicons currently exist. The chunker Comainu [3] (Kozawa et al., 2014) can produce LUW and Bunsetsu based on MeCab outputs. SUW, LUW, and Bunsetsu are annotated in the Balanced Corpus of Contemporary Written Japanese (hereafter BCCWJ) (Maekawa et al., 2014).

The column 'TTR' represents the type-token ratio of the units for the BCCWJ. The TTR of Character is $0.00004 = 7,622/195,898,039$; the TTR of SUW is $0.00176 = 185,136/104,612,418$; the TTR of LUW is $0.02922 = 2,434,721/83,308,386$; and the TTR of Bunsetsu is $0.22221 = 9,485,940/42,688,154$. The large TTR causes modelling difficulty for word embeddings. Therefore, Japanese natural language processing uses word embeddings based on SUW (Asahara, 2018) or characters.

2.3 History of UD Japanese word delimitation

UD Japanese KTC (Tanaka et al., 2016) is the UD corpus based on the Kyoto Corpus. The corpus was resegmented into LUW-like word units and a manually annotated phrase structure tree. The phrase-structure tree was then converted into the UD version 1 standard. However, the maintenance of the UD Japanese KTC stopped after the UD version 2.0 standard.

UD Japanese GSD and PUD are original products by Google (McDonald et al., 2013) and were maintained until version 1.4. The UD Japanese team have maintained them from v2.0 (Tanaka et al., 2016). The word delimitation of v2.0-v2.5 was produced by IBM word segmenter (Kanayama et al., 2000) and manually fixed. Those of v2.6-v2.8 treebanks were based on manual annotation of SUW.

UD Japanese BCCWJ is based on the BCCWJ. As mentioned earlier, the BCCWJ has three delimitations: SUW, LUW, and Bunsetsu. Currently, we only use SUW for word delimitation of the UD Japanese BCCWJ (Omura and Asahara, 2018).

[1] https://ccd.ninjal.ac.jp/unidic/en/
[2] https://taku910.github.io/mecab/
[3] https://github.com/skozawa/Comainu

3 LUW-based UD Japanese

		Sentences	Bunsetsus/LUW	Words/LUW	Bunsetsus/SUW	Words/SUW
BCCWJ	train	40,801	308,648	715,759	308,679	908,738
	dev	8,427	60,697	145,398	60,722	178,306
	test	7,881	56,332	134,475	56,350	166,859
GSD	train	7,050	57,174	130,298	57,357	168,333
	dev	507	4,186	9,531	4203	12,287
	test	543	4,568	10,429	4,588	13,034
PUD	test only	1,000	9,971	22,910	10,008	28,788

Table 2: Basic statistics of the LUW-based word delimitation UD.

We developed an LUW-based UD Japanese corpus based on the UD Japanese BCCWJ, GSD, and PUD. Table 2 shows the basic statistics of SUW, LUW, and Bunsetsu in these treebanks. As can be seen in Table 2, the number of LUW words is small because it contains SUWs. Although LUW delimitation is used to define constituents in the Bunsetsu in the previous section, the number of Bunsetsu also declines because multi-word functional expressions are one LUW.

UD Japanese GSD and PUD are annotated with SUW-based word delimitation, UniDic POS information (XPOS), and Bunsetsu-based dependency relations. Version 2.8 of these treebanks were developed using the conversion rules from the Bunsetsu-based dependency structure, which was originally used in the UD Japanese BCCWJ (Omura and Asahara, 2018).

We manually annotated LUW-based word delimitation, POS, and LEMMA for the UD Japanese GSD and PUD. When we found a discrepancy between SUW and LUW, we modified the SUW-based annotations. The conversion rules adopted both SUW and LUW POS and morphological features. The original data before the conversion are available in the Github repository [4]. Note that the BCCWJ initially has LUW-based word delimitation, POS, and LEMMA information.

4 Experimental Settings

We performed experiments to evaluate the reproducibility of versions 2.5, 2.8, and LUW of the UD Japanese GSD with publicly available language resources: **v2.5 (IBM)** is IBM-word-segmenter-based word delimitation; **v2.8 (SUW)** is SUW word delimitation; **LUW** is LUW word delimitation. The data are split into train, dev, and test. We use train for the training, dev for parameter tuning, and test for evaluation.

The evaluation is performed in three layers for each setting. The first layer is the evaluation of all analysers, whose inputs are raw sentences. The second layer is the evaluation of POS tagging and dependency analysers, whose input is word-delimited sentences (**Gold**). The third layer is the evaluation of dependency analysers, whose input is gold word delimited and POS tagged sentences (**Gold**).

We used UDPipe (Straka and Straková, 2017) as trainable pipeline analysers for tokenisation, tagging, lemmatisation, and dependency parsing. We retrained the UDPipe model with the three-word delimitation of v2.5, v2.8, and LUW corpus (**UDPipe (T)** and **Train**). We used UDpipe v1.2.0 [5]. Since the UDPipe model provided by the LINDAT/CLARIN infrastructure [6] was initially trained by v2.5, we also included the result of the original UDPipe model (**UDPipe (O)** and **Original**). The training of dependency analysis layers of UDPipe can use externally trained word embeddings. We compared the results with and without SUW-based word embeddings, NWJC2vec (Asahara, 2018) (**Train w/o vec** or **Train w/ vec**). SUW and LUW can be reproduced by the morphological analysers MeCab and chunker Comainu, which are trained on data other than UD Japanese. **MeCab** means that we use MeCab-0.996

[4] https://github.com/masayu-a/UD_Japanese-GSDPUD-CaboCha

[5] https://ufal.mff.cuni.cz/udpipe/1

[6] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3131

with UniDic-2.1.2 output for the tokenisation. **Comainu** means that we use Comainu-0.72 output for the tokenisation. [7]

We used evaluation scripts of CoNLL 2018 shared tasks (Zeman et al., 2018). **Words**, **UPOS**, **XPOS**, and **Lemma** are their F_1 scores. **UAS** (Unlabelled Attachment Score) and **LAS** (Labelled Attachment Score) are standard evaluation metrics in dependency parsing results. **CLAS** (Nivre and Fang, 2017) is defined as the labelled F_1-score over all relations except functional and punctuation relations based on **LAS**. **MLAS** (Zeman et al., 2018) is an extension of CLAS, in which function words are not ignored, but treated as features of content words. In addition, the part-of-speech tags and morphological features are evaluated. **BLEX** (Zeman et al., 2018) is another extension of CLAS, incorporating lemmatisation instead of morphological features.

5 Results

Table 3 shows the results. First, we show the reproducibility of word delimitation. Whereas the word delimitation of v2.5 (IBM) was 91.94-91.96% by UDPipe, the values of v2.8 (SUW) and LUW were 96.14% and 95.02%, respectively. SUW and LUW thus are significantly more reproducible than v2.5. When we used MeCab for SUW and Comainu for LUW, the word delimitation accuracies were 96.84% and 97.19%, respectively.

Second, we confirmed the results of UPOS and XPOS. When we used Gold word segmentation as the input, the accuracies of UPOS and XPOS were 96.82-97.39% and 96.34-96.70%, respectively. When we used raw sentences as the input for UDPipe, v2.5 UPOS and XPOS accuracies were 89.3% and 88.98%, respectively, because of their low word-delimitation accuracy. The UPOS and XPOS accuracies of the SUW were 93.96% and 93.29%, respectively. The UPOS and XPOS accuracies of LUW were 92.37% and 92.16%, respectively. When we used MeCab for the tokeniser, the UPOS and XPOS, accuracies of SUW by UD Pipe are 94.42% and 93.58%, respectively. When we used Comainu for the tokeniser, UPOS and XPOS accuracies of LUW by UD Pipe remained at 94.34% and 94.18%, respectively.

Third, the result of lemmatisation shows the disadvantage of LUW. When we used Gold word segmentation as the input, v2.5 and SUW showed 98.93-99.20% LEMMA accuracy. However, LUW showed 93.78%. When we used raw sentences as the input for UDPipe, the LEMMA accuracy of LUW was 89.74%. This is because the lemmatisation of compound morphemes in Japanese is not straightforward. We need other lemmatisation modules for LUW word lemmatisation. When we used Comainu for word delimitation, the LEMMA accuracy of LUW by UD Pipe was 91.32% despite the high tokenisation accuracy (97.19%).

Next, we discuss dependency analysis accuracy. When we used Gold word delimitation, v2.5 outperformed the others. However, because the word delimitation accuracy of v2.5 was low, the UAS and LAS scores dropped from 93.36-95.20% to 75.43-77.91% for the raw sentence. The UAS and LAS scores of SUW were 85.22% and 83.50% with UDPipe, and 88.22% and 86.32% with MeCab for the raw sentences. The UAS and LAS scores of LUW were 83.49% and 82.07% with UDPipe, and 88.16% and 86.45% with Comainu for the raw sentences. When using publicly available word segmenters (MeCab and Comainu), the difference between SUW and LUW for dependency analysis accuracy (UAS, LAS) was not significant. Whereas the CLAS and MLAS results are similar to the UAS and LAS results, the BLEX of LUW is significantly lower than that of SUW. This is because the lemmatisation of LUW is quite difficult. Despite the low lemmatisation accuracy of LUW, the BLEX of LUW outperforms that of v2.5.

[7] These tools can output the XPOS; however, these experiments have ignored inconsistent results.

Treebank	Tokenisation	POS Tagging	Dep. Analysis	Words	UPOS	XPOS	Lemmas	UAS	LAS	CLAS	MLAS	BLEX
v2.5 (IBM)	UDPipe (O)	UDPipe (O)	Original	91.96%	89.35%	88.98%	91.19%	77.91%	76.45%	66.50%	63.94%	65.98%
	Gold	UDPipe (O)	Original	-	96.93%	96.41%	99.07%	92.70%	90.65%	83.55%	80.64%	82.84%
	Gold	Gold	Original	-	-	-	-	94.85%	93.77%	87.08%	86.98%	87.08%
	UDPipe (T)	UDPipe (T)	Train w/o vec	91.94%	89.30%	89.00%	91.14%	77.11%	75.43%	64.84%	62.69%	64.42%
	Gold	UDPipe (T)	Train w/o vec	-	96.82%	96.34%	98.93%	92.12%	89.82%	82.10%	79.19%	81.37%
	Gold	Gold	Train w/o vec	-	-	-	-	94.68%	93.36%	86.29%	86.10%	86.29%
	UDPipe (T)	UDPipe (T)	Train w/ vec	91.94%	89.30%	89.00%	91.14%	77.31%	75.87%	65.71%	63.63%	65.33%
	Gold	UDPipe (T)	Train w/ vec	-	96.82%	96.34%	98.93%	92.58%	90.58%	83.58%	80.76%	82.84%
	Gold	Gold	Train w/ vec	-	-	-	-	95.20%	94.15%	87.85%	87.71%	87.85%
v2.8 (SUW)	MeCab	UD Pipe (T)	Train w/o vec	96.84%	94.42%	93.58%	96.07%	87.38%	85.40%	77.57%	75.04%	77.14%
	UDPipe (T)	UDPipe (T)	Train w/o vec	96.14%	93.96%	93.29%	95.39%	84.40%	82.58%	75.20%	72.81%	74.79%
	Gold	UDPipe (T)	Train w/o vec	-	97.39%	96.52%	99.20%	91.20%	89.10%	83.10%	80.32%	82.64%
	Gold	Gold	Train w/o vec	-	-	-	-	92.28%	91.12%	85.54%	85.00%	85.54%
	MeCab	UDPipe (T)	Train w/ vec	96.84%	94.42%	93.58%	96.07%	88.22%	86.32%	79.28%	76.74%	78.83%
	UDPipe (T)	UDPipe (T)	Train w/ vec	96.14%	93.96%	93.29%	95.39%	85.22%	83.50%	76.85%	74.48%	76.44%
	Gold	UDPipe (T)	Train w/ vec	-	97.39%	96.52%	99.20%	92.05%	90.07%	84.90%	82.13%	84.40%
	Gold	Gold	Train w/ vec	-	-	-	-	93.95%	93.32%	89.07%	88.67%	89.07%
LUW	Comainu	UDPipe (T)	Train w/o vec	97.19%	94.34%	94.18%	91.32%	87.91%	86.16%	78.10%	74.19%	71.78%
	UDPipe (T)	UDPipe (T)	Train w/o vec	95.02%	92.37%	92.16%	89.74%	83.25%	81.83%	72.31%	68.54%	67.10%
	Gold	UDPipe (T)	Train w/o vec	-	96.90%	96.70%	93.78%	92.82%	90.93%	82.66%	78.51%	75.66%
	Gold	Gold	Train w/o vec	-	-	-	-	93.86%	93.23%	85.77%	85.33%	85.77%
	Comainu	UD Pipe (T)	Train w/ vec	97.19%	94.34%	94.18%	91.32%	88.16%	86.45%	78.69%	75.05%	72.52%
	UDPipe (T)	UDPipe (T)	Train w/ vec	95.02%	92.37%	92.16%	89.74%	83.49%	82.07%	72.85%	69.15%	67.69%
	Gold	UDPipe (T)	Train w/ vec	-	96.90%	96.70%	93.78%	93.18%	91.26%	83.27%	79.30%	76.43%
	Gold	Gold	Train w/ vec	-	-	-	-	94.03%	93.74%	87.19%	86.86%	87.19%

Table 3: Results: Reproducibility of versions 2.5, 2.8, and LUW of UD Japanese GSD

Treebank	Tokenisation	UAS			LAS		
		w/o Vec	w/vec	Diff	w/o Vec	w/vec	Diff
v2.5(IBM)	UDPipe	77.11%	77.31%	+0.20	75.43%	75.87%	+0.44
v2.8(SUW)	UDPipe	84.40%	85.22%	+0.82	82.58%	83.50%	+0.92
v2.8(SUW)	MeCab	87.38%	88.22%	+0.84	85.40%	86.32%	+0.92
LUW	UDPipe	83.25%	83.49%	+0.24	81.83%	82.07%	+0.24
LUW	Comainu	87.91%	88.16%	+0.25	86.16%	86.45%	+0.29

Table 4: Effect of Word Embeddings (Subset of Table 3)

Finally, we confirmed the effect of word embeddings for UDPipe. Table 4 shows the effect of word embeddings. The word embeddings NWJC2vec (Asahara, 2018) is based on SUW. Thus, whereas the dependency accuracy of IBM and LUW increased by 0.20-0.44 and 0.24-0.29, respectively, the dependency accuracy of SUW increased by 0.82-0.92. The results suggest that the availability of word embeddings is another important factor in the development of UD language resources. As shown by the presented token-type ratios in Table 1, LUW-based word embeddings are not practical in the current state of Japanese natural language processing. Even though the SUW word embeddings are a subset of LUW word definitions, the dependency accuracy of LUW is comparable to that of SUW.

6 Conclusions

This article presented word delimitation issues in UD Japanese. We provided an overview of the word delimitation standards and the history of UD Japanese, and then developed LUW-based UD Japanese language resources that adopt the word unit as a *syntactic word* in Japanese. We evaluated the reproducibility of several versions of UD Japanese with publicly available resources. The results show that LUW-based UD Japanese is as reproducible as SUW-based UD Japanese, even though LUW-based word embeddings are not available. Lemmatisation of LUWs is still difficult because of their compound morphological structures.

Annotation of the *syntactic word*-based dependency treebank is a difficult task for morphologically rich languages such as Japanese without word delimitation. It took great effort to define morphemes, POSs, compound word constructions, and dependency structures. The work took more than eight years to complete and was finished in 2021. The data were released as version 2.9 of UD Japanese GSDLUW, UD Japanese PUDLUW, and UD Japanese BCCWJLUW.

Our future work will be to adjust the differences in opinions in Japanese natural language processing communities for word delimitation issues. We are also planning to adjust the difference in word delimitation among East Asian languages, such as Chinese and Korean.

Acknowledgements

This work was supported by JSPS KAKENHI (Grant Number JP17H00917) and is a project of the Center for Corpus Development, NINJAL.

References

Masayuki Asahara. 2018. NWJC2Vec: Word embedding dataset from ‘ NINJAL Web Japanese Corpus ’ . *Terminology: International Journal of Theoretical and Applied Issues in Specialized Communication*, 24:7–22, January.

Jayeol Chun, Na-Rae Han, Jena D. Hwang, and Jinho D. Choi. 2018. Building Universal Dependency treebanks in Korean. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2194–2202. European Language Resources Association (ELRA), May.

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

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)*, pages 1019–1024. European Language Resources Association (ELRA), May.

Hiroshi Kanayama, Kentaro Torisawa, Yutaka Mitsuishi, and Jun'ichi Tsujii. 2000. A hybrid Japanese parser with hand-crafted grammar and statistics. In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*, pages 411—-417, July.

Daisuke Kawahara and Sadao Kurohashi. 2006. A fully-lexicalized probabilistic model for Japanese syntactic and case structure analysis. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 176–183. Association for Computational Linguistics, June.

Shunsuke Kozawa, Kiyotaka Uchimoto, and Yasuharu Den. 2014. Adaptation of Long-Unit-Word analysis system to different part-of-speech tagset [in Japanese]. *Journal of Natural Language Processing*, 21(2):379–401.

Taku Kudo and Yuji Matsumoto. 2002. Japanese dependency analysis using cascaded chunking. In *Proceedings of the 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*, pages 1—7. Association for Computational Linguistics, August.

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. Association for Computational Linguistics, July.

Herman Leung, Rafaël Poiret, Tak-sum Wong, Xinying Chen, Kim Gerdes, and John Lee. 2016. Developing Universal Dependencies for Mandarin Chinese. In *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*, pages 20–29. The COLING 2016 Organizing Committee, December.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguti, Makiro Tanaka, and Yasuharu Den. 2014. Balanced corpus of contemporary written Japanese. *Language resources and evaluation*, 48(2):345–371, December.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97. Association for Computational Linguistics, August.

Yugo Murawaki. 2019. On the definition of Japanese word, June. arXiv: 1906.09719 [cs.CL].

Joakim Nivre and Chiao-Ting Fang. 2017. Universal Dependency evaluation. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 86–95. Association for Computational Linguistics, May.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666. European Language Resources Association (ELRA), May.

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. Association for Computational Linguistics, November.

Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing ud 2.0 with UDpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99. Association for Computational Linguistics, August.

Takaaki Tanaka, Yusuke Miyao, Masayuki Asahara, Sumire Uematsu, Hiroshi Kanayama, Shinsuke Mori, and Yuji Matsumoto. 2016. Universal dependencies for japanese. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1651–1658. European Language Resources Association (ELRA), May.

Fei Xia. 2000. The segmentation guidelines for the Penn Chinese Treebank (3.0). *University of Pennsylvania*, November.

Koichi Yasuoka. 2019. Universal Dependencies Treebank of the Four Books in Classical Chinese. In *DADH2019: 10th International Conference of Digital Archives and Digital Humanities*, pages 20–28. Digital Archives and Digital Humanities, December.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21. Association for Computational Linguistics, October.