Introducing Morphology in Universal Dependencies Japanese

Chihiro Taguchi and **David Chiang**
University of Notre Dame, USA
{ctaguchi,dchiang}@nd.edu

Abstract

This paper discusses the need for including morphological features in Japanese Universal Dependencies (UD). In the current version (v2.11) of the Japanese UD treebanks, sentences are tokenized at the morpheme level, and almost no morphological feature annotation is used. However, Japanese is not an isolating language that lacks morphological inflection but is an agglutinative language. Given this situation, we introduce a tentative scheme for retokenization and morphological feature annotation for Japanese UD. Then, we measure and compare the morphological complexity of Japanese with other languages to demonstrate that the proposed tokenizations show similarities to synthetic languages reflecting the linguistic typology.

1 Introduction

This paper introduces morphology-aware tokenization and morphological features to Universal Dependencies (UD) treebanks for Japanese. Since its inception in 2015, the UD project has been developed to cover more than 130 languages as of v2.11 (de Marneffe et al., 2021; Zeman et al., 2022). Its crosslinguistically consistent syntactic and morphological annotation has enabled corpus-based multilingual NLP at a greater scale (Nivre et al., 2020). However, the Japanese treebanks in the current UD v2.11 have divergent policies in terms of tokenization and morphological feature annotation. Specifically, sentences are tokenized by morpheme boundaries with almost no morphological feature assigned, despite the linguistic fact that Japanese has morphological inflection. Given this issue, this paper will propose new tentative schemes for tokenization and morphological annotation that takes into account the synthetic nature of Japanese. Then, we will demonstrate that the retokenized Japanese UD treebanks with these schemes have morphological complexities similar to other synthetic languages. These results agree with the typology of Japanese as a synthetic agglutinative language.

2 Background

This section overviews the Japanese language and the annotation issues that the current Japanese UD treebanks have. It is a typical head-final language with synthetic morphology, where grammatical information is mostly expressed by means of agglutination.

2.1 Orthography

Modern Japanese orthography uses three writing systems: *hiragana* (ひらがな), *katakana* (カタカナ), and *kanji* (漢字). The first two are phonographic writing systems, where each character represents a mora.[1] *Kanji* is a logographic system borrowed from Chinese, and one character may be associated with more than one pronunciation. These three writing systems are used in a mixed manner, where *kanji* is typically used for content words including Chinese loanwords, *katakana* mainly for non-Sino-Japanese loanwords such as from English, and *hiragana* elsewhere. In addition, Japanese orthography does not mark word boundaries, unlike many other orthographies that use spaces for indicating boundaries. These orthographic conventions give rise to various controversies in terms of tokenization and standardized lemmatization.

2.2 Morphology

While Japanese morphology is primarily agglutinative, there is also a limited degree of fusional morphology, where one inflectional morpheme is

[1] A mora is a prosodic unit. A single mora includes a Consonant–Vowel (CV) pair, a single vowel, syllable-final /n/, the last part of a long vowel, and the first part of a geminate consonant. For example, the word *kittinkauntaa* "countertop" consists of eight morae (*ki-t-ti-n-ka-u-n-ta-a*).

Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023), pages 65 - 72

FORM	LEMMA	FEATS
irassyara	*irassyaru*	_
nakat	*nai*	Polarity=Neg
ta	*ta*	_
irassyaranakatta	*irassyaru*	Polarity=Neg|Polite=Resp|Tense=Past|VerbForm=Fin

Table 1: Tokenization, lemmatization, and morphological feature description for (1) with a simplified ConLL-U format. The upper three rows represent the style of the current Japanese UD treebank, and the last row represents the style proposed in this paper. Word forms and lemmas are romanized for readers' convenience.

responsible for more than one feature. For example, the single morpheme *irassyara* in sentence (1) has a grammatical feature of respectful politeness as well as the lexical meaning. Tokenization by a chunk that includes all of enclitics and affixes in a token is called 文節 (*bunsetu*; "sentence parts") in Japanese linguistics.

(1) いらっしゃらなかった
irassyara-nakat-ta
come.RESP-NEG-PST
'(The one respected by the speaker) did not come.'

2.3 Universal Dependencies treebanks

In UD v2.11, Japanese is the second largest language, with approximately 2,849k tokens in total. The seemingly large size is a result of the corpora containing two versions with the same sentences and two different tokenization schemes: Short Unit Word (SUW) and Long Unit Word (LUW). Tokens in SUW are the smallest meaningful units, while LUW's tokenization takes into account compound tokens such as compound nouns and light verb constructions.[2] SUW and LUW largely overlap the notion of *tango* (単語; "word") in the Japanese grammar analyzed by Shinkichi Hashimoto, which is generally taught in the Japanese language education in Japan ("Hashimoto Grammar" (HG) henceforth).

Compared to other treebanks in UD, annotation in Japanese UD is unique in three aspects. First, the tokenization splits at the morpheme level (see the upper three rows of Table 1 for example). This stands in clear contrast with other agglutinative languages in UD, where suffixes are commonly included in one token together with the word root, with their morphological functions expressed as features.

Second, morphological features (FEATS) in Japanese UD treebanks are mostly left blank except for very limited cases such as Polarity=Neg. Other morphemes carrying grammatical features are not provided with any information in FEATS; for instance, grammatical information for RESP and PST in the gloss (1) is not specified as features in Japanese UD (see Table 1).

Third, in the architecture of UD, this strictly morpheme-level tokenization in both SUW and LUW faces a crucial problem: the word form cannot be computed from its lemma and features. For example, although the first token in Table 1 *irassyara* is different from its lemma *irassyaru*, the annotation does not tell us why they have different forms. HG calls the first form *mizenkei* (未然形; "irrealis form"), but this form is not responsible for any specific meaning by itself and therefore is not a morphological feature. Therefore, SUW and LUW fail to capture the morphology of Japanese.

3 Related Work

In the NLP literature on Japanese, the term "morphological analysis" has been used to refer to the task of morphological segmentation, given the fact that the Japanese orthography does not explicitly contain word boundaries (Den et al., 2008; Kudo et al., 2004; Neubig et al., 2011). Since there is no solid linguistic criterion to define what a word is, the smallest meaningful unit (i.e., morpheme) is a stable candidate for tokenizing a language with no orthographic word boundary. This tokenization policy is common in Japanese corpora, as is comprehensively defined in the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Ogura et al., 2011) as SUW and LUW. Existing Japanese morphological analyzers such as MeCab[3] and Sudachi[4] are based on the same policy, and their main concern has chiefly been morpheme-level tokenization and POS tagging while leaving the analysis of morphological features untouched.

[2] For a comprehensive definition and examples, see https://clrd.ninjal.ac.jp/bccwj/en/morphology.html.

[3] http://taku910.github.io/mecab/

[4] https://www.mlab.im.dendai.ac.jp/~yamada/ir/MorphologicalAnalyzer/Sudachi.html

The above-mentioned issues of Japanese UD have already been raised by multiple researchers. Pringle (2016) gives a comprehensive overview of the tokenization of Japanese UD from the viewpoint of general linguistics, concluding that the current tokenization scheme is an artifact of decisions made by the corpora on which the UD Japanese treebanks were based—decisions which UD Japanese should revisit for the sake of the crosslinguistic nature of UD. Murawaki (2019) provides discussion on defining a word in Japanese for UD, and demonstrates that a word (FORM) in Japanese UD does not follow UD's general annotation guideline, which states that "morphological features are encoded as properties of words and there is no attempt at segmenting words into morphemes."[5] However, no actual implementation for retokenization has been realized.

This situation in fact prevents Japanese from being included in crosslinguistic studies with UD data. Çöltekin and Rama (2022) investigate various measures of morphological complexity with more than 50 UD treebanks, but they had to exclude Japanese and Korean treebanks because "no linguistically interesting features were marked despite the fact that both languages are morphologically complex."

4 Retokenization

Given the current issues with Japanese UD, this section proposes tentative alternative annotation schemes that take into account synthetic aspects of Japanese morphology.

4.1 Policies

To define a token in Japanese, we prepared two levels of tokenization policies that reflect Japanese morphological inflection differently. At the first level, each verb and its inflectional morphemes are joined into a single token, which is annotated with appropriate features. These morphemes correspond to 助動詞 (*zyodousi*; "auxiliary verbs") in HG's terms as well as in XPOS of conventional Japanese UD treebanks (see Table 6 in Appendix for details). The last row of Table 1 shows an example retokenized on this level for sentence (1).

The second level also joins verbs and their inflectional morphemes as at the first level; in addition, each noun and its case markings are joined into a single token, which is annotated with appropriate features. These case markings are called 格助詞 (*kakuzyosi*; case particles) in HG's terms; see Table 7 for details. Most of the other types of particles are treated as independent tokens.

The motivation to treat verbal inflection suffixes and case markings at different levels of tokenization is that the morphosyntactic distribution of case markers is freer than those in other agglutinative languages that consider them as part of their morphology. Although the Japanese case markings are functionally similar to case suffixes, their less synthetic distribution is as independent as enclitics and more detached than suffixes (Miyaoka, 2002). For example, Japanese cases always have regular forms and can be attached to material already containing a clitic, whereas affixal morphology tends to have irregular inflection and more limited morphosyntactic distribution. However, as Haspelmath (2015) pointed out, there have been no crosslinguistically viable criteria that distinguish a clitic from an affix. For this reason, we leave the rigid morphosyntactic treatment of Japanese case-marking on hold and instead prepare two levels of schemes corresponding to both of the treatments. Table 2 illustrates the comparison of SUW, LUW, *bunsetu*, and the proposed tokenization schemes.

4.2 Implementation

Since this paper cannot give a decisive answer as to which level is linguistically more suitable to UD, we implemented retokenizers for both of these policies. The retokenization and feature assignment were done fully automatically with rule-based token rejoining, thanks to the fine-grained XPOS annotation in UD Japanese treebanks.[6] We converted the UD_Japanese-GSD and UD_Japanese-GSDLUW treebanks with respect to the two tokenization levels. GSD and GSDLUW are SUW-based and LUW-based treebanks with the same sentences, respectively.

5 Morphological Complexity of Japanese

To confirm the validity of the morphology-aware Japanese UD treebanks, this section reports experiments to measure the morphological complexity of Japanese, which Çöltekin and Rama (2022) could not compare due to the lack of morphological information in current Japanese UD.

[5] https://universaldependencies.org/u/overview/tokenization.html

[6] The codes used in the retokenization process are available here: https://github.com/ctaguchi/ud_ja_standardize.

SUW	魚 *sakana* fish NOUN	フライ *hurai* fry NOUN	を *wo* ACC ADP	食べ *tabe* eat VERB	た *ta* PST AUX	か *ka* Q PART	も *mo* also ADP	しれ *sire* know VERB	ない *nai* NEG AUX	ペルシャ *perusya* Persia NOUN	猫 *neko* cat NOUN
LUW	魚フライ *sakanahurai* fried_fish NOUN		を *wo* ACC ADP	食べ *tabe* eat VERB	た *ta* PST AUX	かもしれない *kamosirenai* may AUX				ペルシャ猫 *perusyaneko* Persian_cat NOUN	
bunsetu	魚フライを *sakanahuraiwo*			食べたかもしれない *tabetakamosirenai*						ペルシャ猫 *perusyaneko*	
proposal (SUW$_1$)	魚 *sakana* fish NOUN _	フライ *hurai* fry NOUN _	を *wo* ACC ADP _	食べた *tabeta* eat.PST VERB Tense=Past VerbForm=Fin		か *ka* Q PART _	も *mo* also ADP _	しれない *sirenai* know.NEG VERB Polarity=Neg Tense=Pres VerbForm=Fin		ペルシャ *perusya* Persia NOUN _	猫 *neko* cat NOUN _
proposal (SUW$_2$)	魚 *sakana* fish NOUN _	フライを *huraiwo* fry.ACC NOUN Case=Acc		食べた *tabeta* eat.PST VERB Tense=Past VerbForm=Fin		か *ka* Q PART _	も *mo* also ADP _	しれない *sirenai* know.NEG VERB Polarity=Neg Tense=Pres VerbForm=Fin		ペルシャ *perusya* Persia NOUN _	猫 *neko* cat NOUN _
proposal (LUW$_1$)	魚フライ *sakanahurai* fried_fish NOUN _		を *wo* ACC ADP _	食べた *tabeta* eat.PST VERB Tense=Past VerbForm=Fin		かもしれない *kamosirenai* may AUX Tense=Pres VerbForm=Fin				ペルシャ猫 *perusyaneko* Persian_cat NOUN _	
proposal (LUW$_2$)	魚フライを *sakanahuraiwo* fried_fish.ACC NOUN Case=Acc			食べた *tabeta* eat.PST VERB Tense=Past VerbForm=Fin		かもしれない *kamosirenai* may AUX Tense=Pres VerbForm=Fin				ペルシャ猫 *perusyaneko* Persian_cat NOUN _	

Table 2: Example of different tokenization schemes (SUW, LUW, bunsetu, and the proposed tokenization) for the sentence 魚フライを食べたかもしれないペルシャ猫 ("A Persian cat that might have eaten fried fish") (Omura and Asahara, 2018). Subscripts on SUW and LUW denote the levels of retokenization proposed in this paper.

5.1 Setup

The measures we used in this study are type–token ratio (TTR), mean size of paradigms (MSP), information in word structure (WS), word entropy (WH), lemma entropy (LH), inflectional synthesis (IS), and morphological feature entropy (MFH) based on the implementation by Çöltekin and Rama (2022). Section D in Appendix illustrates the details of these measures. We compared our retokenized versions of the Japanese GSD and GSD-LUW treebanks with all the treebanks used in their work. For each treebank, we picked 10 samples of 20,000 tokens and averaged the obtained values over the number of samples. Since Japanese orthography is highly logographic (Sproat and Gutkin, 2021), tokens and lemmas are romanized before computation so that orthographic discrepancies among *hiragana*, *katakana*, and *kanji* are ignored.

5.2 Results

Table 3 summarizes the results for selected treebanks.[7] To compare typological differences, the table demonstrates Japanese treebanks (GSD, GSD-LUW, and their retokenized versions), Vietnamese (analytic), English (weakly analytic), Russian (fusional), and Turkish (agglutinative). For Japanese treebanks, there are overall tendencies where LUW, which treats compound nouns and light verb constructions as one token, is more morphologically complex than SUW. In addition, it is evident that including verbal conjugation and nominal

[7]The codes and full results are published in the forked repository: `https://github.com/ctaguchi/mcomplexity`.

Language	Typology	Treebank	TTR	MSP	WS	WH	LH	IS	MFH
Japanese	agglutinative	GSD	0.259	1.075	0.318	9.397	9.192	0.0	1.325
		GSD_1	0.263	1.109	0.365	9.600	9.265	9.8	2.583
		GSD_2	0.400	1.471	0.505	11.242	10.241	11.2	3.030
		GSDLUW	0.320	1.082	0.351	9.433	9.223	0.0	1.296
		$GSDLUW_1$	0.338	1.061	0.448	9.600	9.455	9.7	2.619
		$GSDLUW_2$	0.426	1.065	0.464	11.296	11.037	11.6	3.144
Vietnamese	analytic	VTB	0.166	1.0	0.374	9.964	9.966	0.0	1.253
English	weakly analytic	LinES, GUM, ParTUT	0.207	1.210	0.365	9.572	9.176	5.733	3.701
Russian	fusional	SynTagRus, GSD	0.464	1.479	0.489	11.582	10.797	11.5	3.596
Turkish	agglutinative	IMST	0.399	2.277	0.573	11.719	10.0215	13	3.589

Table 3: Comparison of morphological complexities for the original and retokenized treebanks of Japanese and other typologically diverse languages. A subscript 1 indicates our first level of retokenization (verbs) and a subscript 2 indicates our second level (verbs and nouns). For each measure, the greater a value is, the more morphologically complex the language is. Values for languages with multiple treebanks are averaged.

	vi	en	ru	tr
GSD	**0.9998**	0.8631	0.6708	0.5843
GSD_1	0.6720	0.9349	**0.9992**	**0.9907**
GSD_2	0.6691	0.9337	**0.9993**	**0.9932**
GSDLUW	**0.9998**	0.8612	0.6689	0.5823
$GSDLUW_1$	0.6823	0.9390	**0.9988**	**0.9877**
$GSDLUW_2$	0.6713	0.9352	**0.9990**	**0.9890**

Table 4: Pearson correlation matrix for the selected languages and Japanese treebanks. A subscript 1 indicates our first level of retokenization (verbs) and a subscript 2 indicates our second level (verbs and nouns).

case-marking in morphological annotation leads to higher complexity.

We also notice numerical similarities between the conventional Japanese treebanks (GSD and GSDLUW) and the Vietnamese treebank. In fact, Pearson's correlation matrix shown in Table 4 numerically demonstrates that the measured morphological complexities of conventional treebanks are the most similar to Vietnamese, an analytic language. In contrast, the retokenized treebanks have the highest similarity scores with Russian followed by Turkish, which are both synthetic languages. It is notable that Russian and Turkish do not show much contrast despite their typological difference in the degree of fusion. This is likely due to the limitation of the morphological complexity measures used in this experiment which take into account the distribution of tokens, lemmas, and morphological features but do not consider how a token is morphologically derived from a lemma. A possible way to measure fusionality is to measure the edit distance between a lemma and a surface form that is weighted more on substitution and deletion so that agglutinative morphology (insertion) would score lower and be distinguished from fusional inflections.

Regarding IS and MFH, which take into account morphological features in their variables, it is notable that (i) the IS score for the conventional Japanese treebanks is 0 while our retokenized treebanks show much higher complexity (9.7–11.6) rather close to synthetic languages, and (ii) the MFH of our retokenized treebanks stands between an analytic language and synthetic languages. These results reflect the typological characteristics of Japanese as an agglutinative language.

6 Concluding Remarks

This paper has argued for morphology-aware tokenization policies for UD Japanese treebanks and conducted an experiment that measures the morphological complexity of Japanese based on the retokenized treebanks with morphological features. In doing so, we proposed new annotation schemes for tokenization and morphological features in Japanese. The results showed that, although the morphological complexity of the current Japanese UD resembled that of an isolating language, our retokenized treebanks have scores more similar to synthetic languages, which reflect the typological reality of Japanese. The proposed tokenization will also be suitable for developing UD treebanks for other Japanese–Ryukyuan languages that syntactically have a similar structure to Japanese but can be morphologically more fusional. Furthermore, tokenization and morphological annotation conforming to UD's general guidelines enable crosslinguistic comparative studies; therefore, discussions for further cross-treebank consistencies are required.

7 Acknowledgments

I thank Dr. Yugo Murawaki and Dr. So Miyagawa for our discussion about the tokenization policy of current Japanese UD. I am also grateful to Dr. Çağrı Çöltekin for giving us advice on reproducing the results on morphological complexity. This material is based upon work supported by the National Science Foundation under Grant No. BCS-2109709. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Çağrı Çöltekin and Taraka Rama. 2022. What do complexity measures measure? correlating and validating corpus-based measures of morphological complexity. *Linguistics Vanguard*.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

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).

Martin Haspelmath. 2015. Defining vs. diagnosing linguistic categories: A case study of clitic phenomena. In Joanna Blaszczak, Dorota Klimek-Jankowska, and Krzysztof Migdalski, editors, *How categorical are categories?: New approaches to the old questions of noun, verb, and adjective*, pages 273–304. De Gruyter Mouton, Berlin, München, Boston.

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.

Osahito Miyaoka. 2002. 語とはなにか：エスキモー語から日本語をみる (*go to wa nani ka: esukimoogo kara nihongo wo miru*) [*What is a word: looking at Japanese from the perspective of Eskimo*]. Sanseido.

Yugo Murawaki. 2019. On the definition of japanese word.

Graham Neubig, Yosuke Nakata, and Shinsuke Mori. 2011. Pointwise prediction for robust, adaptable Japanese morphological analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 529–533, Portland, Oregon, USA. Association for Computational Linguistics.

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.

Hideki Ogura, Hanae Koiso, Yumi Fujiike, Sayaka Miyauchi, Hikari Konishi, and Yutaka Hara. 2011. 現代日本語書き言葉均衡コーパス：形態論情報規程集第4版（下） (*Gendai nihongo kakikotoba kinkou koopasu: keitairon zyouhou kiteisyuu dai 4 han* (*ge*)) [*Balanced Corpus of Contemporary Written Japanese: rules on morphological information Ver. 4* (*Vol. 2*)]. 国立国語研究所内部報告書 [Kokuritu kokugo kenkyuuzyo naibu houkokusyo]. National Institute for Japanese Language and Linguistics.

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.

Gregory Pringle. 2016. Thoughts on the Universal Dependencies proposal for Japanese: The problem of the word as a linguistic unit. `http://www.cjvlang.com/Spicks/udjapanese.html`, Date accessed: November 10, 2022.

Richard Sproat and Alexander Gutkin. 2021. The taxonomy of writing systems: How to measure how logographic a system is. *Computational Linguistics*, 47(3):477–528.

Daniel Zeman et al. 2022. Universal dependencies 2.11. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

A Glossing Abbreviations

CAU — causative; COP — copula; DAT — dative; IN — inessive; NEG — negative; NMLZ — nominalizer; PASS — passive; PAST — past; POL — polite; PRES — present (non-past); Q — interrogative particle; RESP — respectful form; TOP — topic.

B Verbal and adjectival inflection in Japanese

Stem form	Verbs Ending	Verbs *kak-* "to write"	*-i* adjectives Ending	*-i* adjectives *naga-* "long"	
Irrealis 未然形	*-a* (*-o*)	*kaka-*, *kako-*	*-karo*	*nagakaro-*	*daro-*
Continuative 連用形	*-i*	*kaki-*, *kai-*	*-ku*, *-kat*	*nagaku-*, *nagakat-*	*de*, *dat-*
Terminal 終止形	*-u*	*kaku*	*-i*	*nagai*	*da*
Attributive 連体形	*-u*	*kaku*	*-i*	*nagai*	*na*
Hypothetical 仮定形	*-e*	*kake-*	*-kere*	*nagakere-*	*nara*
Imperative 命令形	*-e*	*kake*	—	—	—

Table 5: A concise conjugation table for Modern Japanese verbs, *-i* adjectives, and copula.

POS	Form	Feature	Formation	Example
VERB	Negative	`Polarity=Neg`	irr. + *-nai*	*kakanai*
	Passive	`Voice=Pass`	irr. + *-(ra)reru*	*kakareru*
	Causative	`Voice=Cau`	irr. + *-(sa)seru*	*kakaseru*
	Volitional	`Mood=Opt`	irr. + *-(y)ou*	*kakou*
	Polite	`Polite=Form`	cont. + *-masu*	*kakimasu*
	Progressive converb (1)	`Aspect=Prog|VerbForm=Conv`	cont. + *-nagara*	*kakinagara*
	Progressive converb (2)	`Aspect=Prog|VerbForm=Conv`	cont. + *-tutu*	*kakitutu*
	Prospective	`Aspect=Prosp`	cont. + *-sou*	*kakisou*
	Exemplification	`VerbForm=Exem`	cont. + *-tari*	*kaitari*
	Past	`Tense=Past`	cont. + *-ta*	*kaita*
	Past conditional	`Mood=Cnd|Tense=Past`	cont. + *-tara*	*kaitara*
	Converb	`VerbForm=Conv`	cont. + *-te*	*kaite*
	Infinitive	`VerbForm=Inf`	cont. + -Ø	*kaki*
	Conditional	`Mood=Cnd`	hyp. + *-ba*	*kakeba*
	Potential	`Mood=Pot`	hyp. + *-ru*	*kakeru*
ADJ	Exemplification	`VerbForm=Exem`	cont. + *-tari*	*nagakattari*
	Past	`Tense=Past`	cont. + *-atta*	*nagakatta*
	Past conditional	`Mood=Cnd|Tense=Past`	cont. + *-attara*	*nagakattara*
	Converb	`VerbForm=Conv`	cont. + *-te*	*nagakute*
	Infinitive	`VerbForm=Inf`	cont. + Ø	*nagaku*
	Conditional	`Mood=Cnd`	hyp. + *-ba*	*nagakereba*

Table 6: Verbal conjugation of Modern Japanese and its correspondence to UD features. Note that `VerbForm=Exem` is a proposed feature that is currently not part of UD features. The abbreviations irr., cont., and hyp. stand for the stem forms (irrealis, continuative, hypothetical, respectively).

C Nominal inflection in Japanese

Case	Feature	Morpheme	*neko* "cat"
Nominative	`Case=Nom`	*-ga*	*neko-ga*
Genitive	`Case=Gen`	*-no*	*neko-no*
Dative	`Case=Dat`	*-ni*	*neko-ni*
Accusative	`Case=Acc`	*-o*	*neko-o*
Lative	`Case=Lat`	*-e*	*neko-e*
Ablative	`Case=Abl`	*-kara*	*neko-kara*
Locative	`Case=Loc`	*-de*	*neko-de*
Comitative	`Case=Com`	*-to*	*neko-to*
Comparative	`Case=Cmp`	*-yori*	*neko-yori*

Table 7: Tentative feature assignment for case particles (*kakuzyosi*; 格助詞).

D Definitions of the measures

The morphological complexity measures by Çöltekin and Rama (2022) are defined as:

$$\mathrm{TTR} := \frac{|\{T\}|}{|T|}$$

$$\mathrm{MSP} := \frac{|\{T\}|}{|\{L\}|}$$

$$\mathrm{WS} := \frac{|T|}{|\mathrm{compress}(T)|} - \frac{|T_{\mathrm{rand}}|}{|\mathrm{compress}(T_{\mathrm{rand}})|}$$

$$\mathrm{WH} := -\sum_i p(t_i) \log p(t_i)$$

$$\mathrm{LH} := -\sum_i p(l_i) \log p(l_i)$$

$$\mathrm{IS} := |\{\Phi\}|$$

$$\mathrm{MFH} := -\sum_i p(\phi_i) \log p(\phi_i),$$

where T is a list of tokens in the sample, $\{\cdot\}$ a set (i.e., without duplication), $|\cdot|$ the length, T_{rand} the sample after randomly changing characters of its tokens, $\mathrm{compress}(\cdot)$ a compression function, $p(t_i)$ the probability of a token type t_i, $p(l_i)$ the probability of a lemma type l_i, Φ a list of features used in verbs, and $p(\phi_i)$ the probability of a feature type ϕ_i. In the actual implementation, `zlib`'s compression function was used for measuring WS.