J-UɴɪMᴏʀᴘʜ: Japanese Morphological Annotation through the Universal Feature Schema

Kosuke Matsuzaki♣ **Masaya Taniguchi**♡ **Kentaro Inui**◇♣♡ **Keisuke Sakaguchi**♣♡

♣Tohoku University   ♡RIKEN   ◇MBZUAI

matsuzaki.kosuke.r7@dc.tohoku.ac.jp

 github.com/cl-tohoku/J-UniMorph

Abstract

We introduce a Japanese Morphology dataset, J-UɴɪMᴏʀᴘʜ, developed based on the UniMorph feature schema. This dataset addresses the unique and rich verb forms characteristic of the language's agglutinative nature. J-UɴɪMᴏʀᴘʜ distinguishes itself from the existing Japanese subset of UniMorph, which is automatically extracted from Wiktionary. On average, the Wiktionary Edition features around 12 inflected forms for each word and is primarily dominated by denominal verbs (i.e., [noun] + *suru* (do-PRS)). Morphologically, this inflection pattern is same as the verb *suru* (do). In contrast, J-UɴɪMᴏʀᴘʜ explores a much broader and more frequently used range of verb forms, offering 118 inflected forms for each word on average. It includes honorifics, a range of politeness levels, and other linguistic nuances, emphasizing the distinctive characteristics of the Japanese language. This paper presents detailed statistics and characteristics of J-UɴɪMᴏʀᴘʜ, comparing it with the Wiktionary Edition. We will release J-UɴɪMᴏʀᴘʜ and its interactive visualizer publicly available, aiming to support cross-linguistic research and various applications.

1 Introduction

Universal Morphology (UniMorph) is a collaborative project that delivers a wide-ranging collection of standardized morphological features for over 170 languages in the world (Sylak-Glassman, 2016; McCarthy et al., 2020). UniMorph feature schema comprises over 212 feature labels across 23 dimensions of meaning labels, such as tense, aspect, and mood. More concretely, UniMorph dataset consists of a lemma coupled with a set of morphological features that correspond to a specific inflected form, as illustrated by the following example:

走る/ *hashi-ru*　走った/ *hashi-tta*　V;PST;IPFV

where the original form (lemma) "*hashi-ru*" (走る, run-PRS) is inflected to "*hashi-tta*" (走った,

run-PST) to indicate the past tense (PST) and imperfective aspect (IPFV) as morphological features.

The challenge of morphological (re)inflection, which started with the SIGMORPHON 2016 Shared Task (Cotterell et al., 2016), involves generating an inflected form from a given form and its corresponding morphological feature. This effort has continued over years, covering multiple shared tasks (Cotterell et al., 2017, 2018; McCarthy et al., 2019; Vylomova et al., 2020; Pimentel et al., 2021; Kodner et al., 2022; Goldman et al., 2023).

The SIGMORPHON–UniMorph 2023 Shared Task 0 (Goldman et al., 2023) released a Japanese Morphology dataset,[1] which was automatically extracted from Wiktionary. This Wiktionary Edition, on average, highlights 12 inflected forms for each word. It mainly consists of denominal verbs, which are formed by combining a noun with a light verb, and their inflection patterns are morphologically same as the verb "*suru*" (do-PRS).

We propose J-UɴɪMᴏʀᴘʜ. It aims to focus on basic verbs found at the N5 level of the Japanese Language Proficiency Test (JLPT), and it excludes denominal verbs with identical inflection patterns. Our aim was to incorporate a diverse range of expression forms, resulting in an average of 118 inflected forms per word. It includes honorifics, varying levels of politeness, and imperatives with fine-grained distinctions, showcasing the distinctive features of the Japanese language. While only a few languages have manually curated UniMorph resources that extend beyond Wiktionary, J-UɴɪMᴏʀᴘʜ has been carefully designed and created, sharing the same motivation as the project for Korean (Jo et al., 2023).

This paper begins with a brief overview of Japanese verbs, detailing the criteria for labeling J-UɴɪMᴏʀᴘʜ (§2). We then explain the data creation process (§3). As illustrated in Figure 1, this

---

[1] https://github.com/sigmorphon/2023InflectionST/

Table A: Conjugation and politeness type

Seed verb	Politeness type	Conj. type
hashiru	Basic	Regular I
au	Basic	Regular I
taberu	Basic	Regular II
meshiagaru	Respectful (*taberu*)	Regular I
itadaku	Humble (*taberu*)	Regular I
:	:	:
wataru	Basic	Regular I

Table B: Inflection suffix

Suffix
-anai
-imasu
-asemasu
:
-emasen

Table C: Modification for extra inf.

R_strip	L_add	R_add
–	–	-desu
-masu	o-	-ninarimasu
-masu	–	-raremashita
:	:	:
–	–	–

Table D: Proposed UniMorph labels

Prefix and suffix	Labels
-anai	V;PRS;IPFV;…
-anai-desu	V;PRS;IPFV;…
-imasu	V;PRS;IPFV;…
o- -ininarimasu	V;FORM;ELEV;…
-asemasu	V;PRS;IPFV;…
-aseraremashita	V;PST;PFV;…
:	:
-emasen	V;PRS;IPFV;…

Table E: Number of google search hits

Inflected Form	Hits
hashir-anai	5,480,000
hashir-anai-desu	497,000
hashir-imasu	22,700,000
o-hashir-i-ninari…	8
hashir-asemasu	1,170,000
hashir-ase-rare…	266,000
:	:
watar-emasen	230,000

Legend: ⊗ Product, ⬚ Copy row, ⊘ Replace, ⊞ Add row, ⊞ Add column

Generate — Verb inflection tool: kamiya-codec

Output

Output
hashir-anai
hashir-imasu
hashir-asemasu
:
watar-emasen

Inflected Form
hashir-anai
hashir-anai-desu
hashir-imasu
o-hashir-i-ninarimasu
hashir-asemasu
hashir-ase-raremashita
:
watar-emasen

Inflected Form	Labels
hashir-anai	V;PRS;…
hashir-anai-desu	V;PRS;…
hashir-imasu	V;PRS;…
o-hashir-i-ninari…	V;FORM;…
hashir-asemasu	V;PRS;…
hashir-ase-rare…	V;PST;…
:	:
watar-emasen	V;PRS;…

Manually annotated

Discard if Hits ≤ 10

J-UniMorph (our dataset)

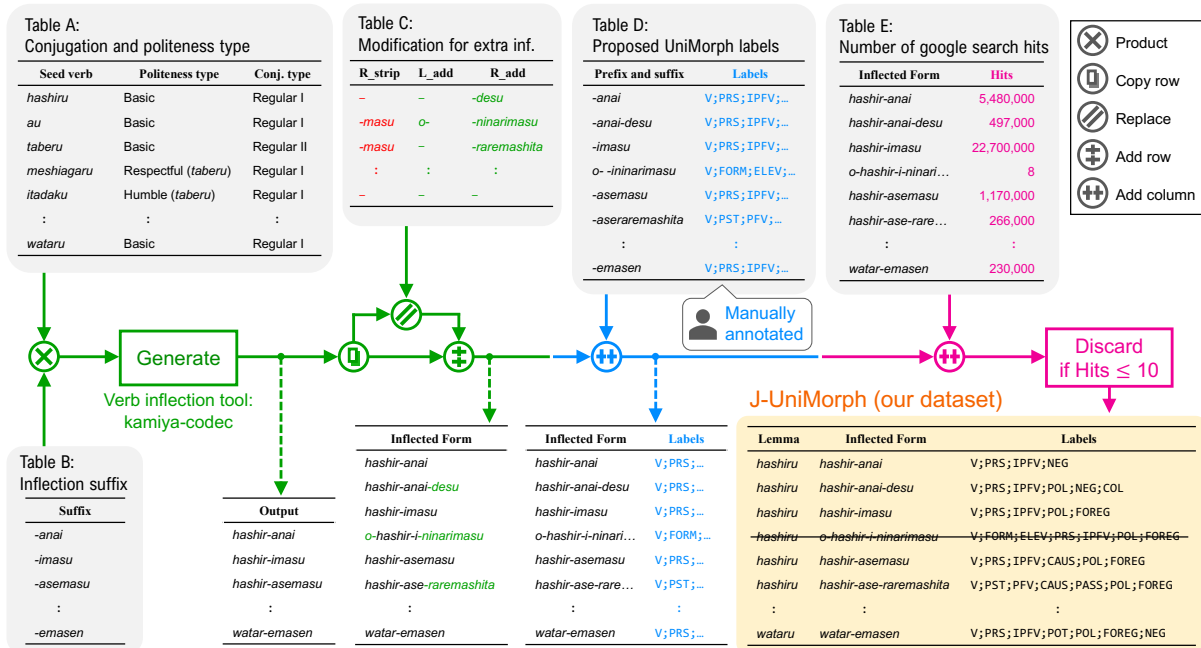| Lemma	Inflected Form	Labels
hashiru	*hashir-anai*	V;PRS;IPFV;NEG
hashiru	*hashir-anai-desu*	V;PRS;IPFV;POL;NEG;COL
hashiru	*hashir-imasu*	V;PRS;IPFV;POL;FOREG
~~*hashiru*~~	~~*o-hashir-i-ninarimasu*~~	~~V;FORM;ELEV;PRS;IPFV;POL;FOREG~~
hashiru	*hashir-asemasu*	V;PRS;IPFV;CAUS;POL;FOREG
hashiru	*hashir-ase-raremashita*	V;PST;PFV;CAUS;PASS;POL;FOREG
:	:	:
wataru	*watar-emasen*	V;PRS;IPFV;POT;POL;FOREG;NEG

Figure 1: Overview of the J-UNIMORPH creation process: First, we generate inflected forms from seed verbs (Table A, detailed in §3.1) and inflection suffix (Table B, detailed in §3.2) using the verb inflection tool, kamiya-codec. This is followed by modifying and adding inflected forms that the tool does not cover (Table C, detailed in §3.2). Second, Japanese native speakers annotate UniMorph labels to each form (Table D, detailed in §2). Finally, we apply a frequency filter to discard infrequent inflected forms (Table E, detailed in §3.3).

process includes three main steps: (1) generating inflected forms (Generation), (2) assigning Uni-Morph labels (Annotation), and (3) removing incorrect or infrequent forms based on frequency (Filtering). Finally, a comparative analysis (§4) between J-UNIMORPH and the Wiktionary Edition shows that J-UNIMORPH includes more commonly used verbs and a wider variety of inflected forms than the Wiktionary Edition, with a slightly larger size (12,687 vs. 12,000).

We have released J-UNIMORPH and its interactive visualizer, aiming to provide a useful resource for cross-linguistic studies, Japanese language learning support, and various applications.

2 Features Schema in J-UNIMORPH

Verbs in Japanese are broadly categorized into three conjugation types: Regular I verbs, Regular II verbs, and Irregular verbs (Kamiya, 2001). Among these, the Irregular verbs include only "*kuru*" (come-PRS) and "*suru*" (do-PRS).[2] Table 1

Regular I verbs (I型動詞，五段活用動詞)
a-u (会う, meet), *ik-u* (行く, go), *kak-u* (書く, write), *kik-u* (聞く, listen), *hashir-u* (走る, run)

Regular II verbs (II型動詞，一段活用動詞)
ki-ru (着る, wear/put on), *kotae-ru* (答える, answer), *tabe-ru* (食べる, eat), *mi-ru* (見る, see/watch)

Table 1: Examples of Regular I and II Verbs

provides examples of Regular I and II verbs.

The authors, who are all native Japanese speakers with Linguistics backgrounds, have carefully and thoroughly discussed to determine the alignment between the inflection patterns and their Uni-Morph feature labels.[3] In this section, we review the common Japanese inflections such as politeness (§2.1), mood including imperatives (§2.2), tense and aspect (§2.3), negation (§2.4), passive (§2.5), and causative (§2.6), and the criteria for labeling J-UNIMORPH. We note that some inflected forms share the same spelling but have ambiguous or multiple meanings, and we annotate these as distinct entries in J-UNIMORPH for clarity.

---

[2] In Japanese, *denominal verbs* are formed by combining a noun with the light verb "*suru*." For example, "*benkyo*" (study-N) becomes "*benkyo-suru*" (study-V;PRS). These verbs share the same inflection pattern as "*suru*" (do-V;PRS). Given their identical inflection pattern, we have excluded denominal verbs from the J-UNIMORPH.

[3] The "label" is also referred to as "tag" recently (McCarthy et al., 2020; Batsuren et al., 2022).

2.1 Politeness

Honorific speech (*Keigo*), which conveys politeness, is primarily classified into three types: polite form (*Teineigo*), respectful form (*Sonkeigo*), and humble form (*Kenjōgo*). We explain the characteristics, usage, and applicable labels in the following.

Polite form (*Teineigo*) Polite form is a form that conveys respect to the reader or listener, and it uses the "*-desu/masu*" form. The level of politeness can be further heightened when used with respectful or humble form (Hirabayashi and Hama, 1988). The UniMorph Schema includes the label POL (Polite), so we assign this label to these form. Additionally, the schema provides the label FOREG (Formal register) for the Japanese "*mas(u)*-style" (Sylak-Glassman, 2016); therefore we have also assigned FOREG to the "*-masu*" form.

Respectful form (*Sonkeigo*) The respectful form of expression elevates the person who should be respected, and is typically used for superiors and customers. This is not used for individuals within the same group or for one's own actions. Most verbs generally take the form of "*-(ra)re-ru*," and "*o—ninaru*," where the verb's inflection occurs between the "*o*" and "*ninaru*." Some verbs also take lexical honorifics, where the word itself changes to express respect, such as changing "*iku*" (go-PRS[4]) to "*irassharu*" (go-PRS;ELEV).

Since these lexical honorifics involve changes beyond simple affixation while maintaining the same part of speech, we treat them as "inflections" of basic verbs. This decision is primarily motivated by their practical use, as they are commonly used in place of basic verbs when expressing respect.

The "*o—ninaru*" form is commonly used for verbs that do not have any lexical honorific. Both the lexical honorific and the "*o—ninaru*" form are labeled with FORM+ELEV (Formal, Referent Elevating), following the UniMorph Schema (Sylak-Glassman, 2016). The "*-(ra)re-ru*" form is assigned only ELEV without FORM. This choice is based on the consideration that this form conveys a lower level of respect compared to the "*o—ninaru*" and the lexical honorific, despite slightly deviating from the schema's definition (Sylak-Glassman, 2016). The following examples illustrate the verb "*iku*" (go-PRS) with a lexical honorific and "*au*" (meet-PRS) without a lexical honorific.

行く/ *iku*
行かれる/ *ika-reru*
V;PRS;IPFV;**ELEV**

行く/ *iku*
いらっしゃる/ *irassharu*
V;**FORM**;**ELEV**;PRS;IPFV

会う/ *au*
会われる/ *awa-reru*
V;PRS;IPFV;**ELEV**

会う/ *au*
お会いになる/ *o-ai-ninaru*
V;**FORM**;**ELEV**;PRS;IPFV

Humble form (*Kenjōgo*) The humble form conveys respect by lowering oneself or one's group in comparison to the person deserving respect. In business contexts, it is used even when referring to the actions of one's own company's superiors, especially when addressing customers. Most verbs mainly take the form of "*o—suru*," where the verb's inflection occurs between the "*o*" and "*suru*." Some verbs also take lexical honorifics. These are labeled as FORM+HUMB (Formal, Speaker Humbling), following the UniMorph Schema (Sylak-Glassman, 2016). The examples below demonstrate the use of the verb "*iku*" (go-PRS) with the lexical honorific and "*kaku*" (write-PRS) without a lexical honorific.

行く/ *iku*
伺う/ *ukagau*
V;**FORM**;**HUMB**;PRS;IPFV

書く/ *kaku*
お書きする/ *o-kaki-suru*
V;**FORM**;**HUMB**;PRS;IPFV

The complexity of Japanese honorifics and their inflection patterns is further complicated by lexical honorifics corresponding to multiple basic forms, and vice versa. For instance, the humble verb "*ukagau*" corresponds to three basic verbs: "*kuru*" (come), "*iku*" (go), and "*kiku*" (ask/listen). On the other hand, the basic verb "*iku*" (go) is associated with three humble verbs: "*mairu*," "*ukagau*," and "*agaru*." In Appendix A, we provide the correspondence between the basic forms and lexical honorifics adopted in J-UNIMORPH.

2.2 Mood

In terms of expressing mood, we deal with the following five categories: Imperative, Intentive, Optative, Potential, and Permissive.

Imperative Japanese has a variety of imperative expressions, as shown in Table 2. This table compiles the inflection and label correspondence of the verb "*tabe-ru*" (eat-PRS) as an example, organizing them into four groups based on the similarity of their label sets. Each group's inflected forms are

---

[4]In the main text, only the relevant label set is presented for brevity.

Inflected form	Romanization	Label
食べろ	*tabe-ro*	V;IMP;OBLIG
食べな	*tabe-na*	V;IMP;OBLIG;COL
食べなさい	*tabe-nasai*	V;IMP;OBLIG;POL
食べて	*tabe-te*	V;IMP;COL
食べてください	*tabe-te-kudasai*	V;IMP;POL
お食べください	*o-tabe-kudasai*	V;FORM;IMP;POL
食べるな	*tabe-ru-na*	V;IMP;OBLIG;NEG
食べないで	*tabe-nai-de*	V;IMP;NEG;COL
食べないでください	*tabe-nai-de-kudasai*	V;IMP;POL;NEG
お食べにならないでください	*o-tabe-ni-naranai-de-kudasai*	V;FORM;IMP;POL;NEG
召し上がれ	*meshiagar-e*	V;FORM;ELEV;IMP;OBLIG
召し上がりな	*meshiagar-i-na*	V;FORM;ELEV;IMP;OBLIG;COL
召し上がりなさい	*meshiagar-i-nasai*	V;FORM;ELEV;IMP;OBLIG;POL
召し上がって	*meshiaga-tte*	V;FORM;ELEV;IMP;COL
召し上がってください	*meshiaga-tte-kudasai*	V;FORM;ELEV;IMP;POL
お召し上がりください	*o-meshiagar-i-kudasai*	V;FORM;ELEV;IMP;POL;COL
召し上がるな	*meshiagar-u-na*	V;FORM;ELEV;IMP;OBLIG;NEG
召し上がらないで	*meshiagar-a-nai-de*	V;FORM;ELEV;IMP;NEG;COL
召し上がらないでください	*meshiagar-a-nai-de-kudasai*	V;FORM;ELEV;IMP;POL;NEG
お召し上がりにならないでください	*o-meshiagar-i-ni-naranai-de-kudasai*	V;FORM;ELEV;IMP;POL;NEG;COL

Table 2: Correspondence between the imperative form and labels, using the verb "*taberu*" (食べる, eat).

roughly sorted by the strength of degree of command, from strong to weak. All forms in Table 2 are labeled IMP (Imperative).

In Table 2, the term "*tabe-ro*" (Do eat!), representing the most forceful command, is annotated with OBLIG (Obligative) due to its compelling nature. This expression is rarely used in everyday conversations as it comes across as overly authoritative. For colloquial forms used in informal speech such as "*tabe-na*" (Eat.), COL (Colloquial) is assigned. For forms that include polite expressions such as "*-nasai*" and "*-kudasai*," POL (Polite) is assigned.

The bottom two groups of Table 2 show imperative inflection patterns and their corresponding labels for lexical honorifics "*meshiagar-u*" (eat-PRS;ELEV), which is one of the respectful forms of the basic verb "*tabe-ru*" (eat-PRS). For these instances, we also assign FORM+ELEV labels (§2.1).

Intentive   Intentive forms such as "*-yō*," "*-ō*," and "*-mashō*" are marked with INTEN (Intentive). Since "*-mashō*" is one of the inflections of the polite form "*-masu*," it is additionally annotated with POL+FOREG (Polite, Formal register) (§2.1). Below are examples of intentive expressions, where these are the inflection of "*tabe-ru*" (eat-PRS).

ピザを食べよう。　　ピザを食べましょう。
Piza-o tabe-yō.　　*Piza-o tabe-mashō*.
Let's eat pizza.　　Let's eat pizza. (Polite)

Optative   Subjective desires are expressed with "*-tai*," and objective ones with "*-tagaru*." We distinguish these two optative expressions with the label OPT (Optative-Desiderative), associated with person specification (1: first person, 3: third person). Below are examples with the verb "*hashir-u*" (run).

走る/ *hashir-u*
走りたい/ *hashir-i-tai*
V;PRS;IPFV;OPT;1
e.g., I want to run. (*Watashi-wa hashir-i-**tai***)

走る/ *hashir-u*
走りたがる/ *hashir-i-tagaru*
V;PRS;IPFV;OPT;3
e.g., He wants to run. (*Kare-wa hashir-i-**tagaru***)

Potential   We assign the label POT (Potential) to expressions that indicate possibility. For Regular I verbs, the suffix "*-eru*" is attached, while Regular II verbs take "*-(ra)reru*," which is identical to the respectful form (§2.1). In J-UNIMORPH, we include these forms as separate entries. Below are examples, with "*kaku*" (write-PRS) being a Regular I verb and "*miru*" (look-PRS) a Regular II verb.

書く/ *kak-u*　　　　見る/ *mi-ru*
書ける/ *kak-eru*　　見られる/ *mi-rareru*
V;PRS;IPFV;**POT**　　V;PRS;IPFV;**POT**

Permissive The expression "*-(sa)se-te-itadaku*" is used to politely request permission, demonstrating humility.[5] We assigned this form with FORM+HUMB+PERM (Formal, Speaker Humbling, Permissive). The following examples demonstrate annotated suffixes for "*-(sa)se-te-itadaki-masu*" with V;**FORM**;**HUMB**;PRS;IPFV;**POL**;**FOREG**;**PERM**.

(a) 私から答えさせていただきます。
*Watashi-kara kotae-**sase-te-itadaki-masu***.
(If allowed,) I will answer (the question).[6]

(b) [店先の貼り紙で] 本日は休ませていただきます.
*Honjitsu-wa yasuma-**se-te-itadaki-masu***.
[Notice at the store front] (Our store) will be closed today. (No specific permission is required)

2.3 Tense and Aspect

There are two forms to express tense or aspect: *ta*-form and *ru*-form. The "*ta*" and "*ru*" respectively represent verb endings such as "*tabe-ta*" (eat-PST) or "*tabe-ru*" (eat-PRS). From a tense perspective, these forms represent the contrast between "past" and "non-past," while from an aspect perspective, they represent the contrast between "perfective" and "imperfective" (Kato and Fukuchi, 1989).

Japanese does not have a distinct form to explicitly distinguish between present and future. Future tense is expressed by adverbial elements such as "next week" or "tomorrow," so we do not assign the label FUT (Future) to the *ru*-form.

Based on the above considerations, the *ta*-form is assigned the label PST+PFV (Past, Perfective), while the *ru*-form is assigned the label PRS+IPFV (Present, Imperfective). The following are examples of the verb "*hashi-ru*" (run-PRS).[7]

走る/ *hashi-ru*	走る/ *hashi-ru*
走る/ *hashi-ru*	走った/ *hashi-tta*
V;PRS;IPFV	V;PST;PFV

Prospective forms such as "*-darō*" and "*-deshō*" are marked with PROSP (Prospective). As "*-deshō*" is one of the inflections of the polite form "*-desu*,"

it is also annotated with POL (Polite). An example of the usage of "*-deshō*" is presented below.

明日は晴れるでしょう。
*Ashita-wa hare-ru-**deshō***.
It will be sunny tomorrow.

2.4 Negation

Negation in Japanese is primarily expressed through the suffixes "*-nai*" or "*-masen*," and in J-UNIMORPH, the label NEG (Negative) is assigned to indicate negation. Since "*-masen*" is an inflection of the polite form "*-masu*," we assign the label POL+FOREG+NEG (Polite, Formal register, Negative) to it. Another polite negation form, "*-nai-desu*", is commonly used in colloquial speech, and thus, the label POL+NEG+COL (Negative, Colloquial) is applied to it.

Importantly, neither "*-nai*" (NEG) nor "*-desu*" (POL) alone conveys a colloquial tone; however, COL becomes apparent when they are combined, highlighting the non-monotonic compositional nature of verb inflection in Japanese. Below are examples of "*mi-ru*" (look-PRS).

見る/ *mi-ru*	見る/ *mi-ru*
見ない/ *mi-nai*	見ないです/ *mi-nai-desu*
V;PRS;IPFV;**NEG**	V;PRS;IPFV;**POL**;**NEG**;**COL**

見る/ *mi-ru*
見ません/ *mi-masen*
V;PRS;IPFV;**POL**;**FOREG**;**NEG**

2.5 Passive

The passive voice (PASS) are expressed through the suffix "*-(ra)re-ru*," which shares the same form as the respectful form (§2.1) and also potential form (§2.2). In J-UNIMORPH, we categorize these forms as distinct entries for clarity. An example of the use of the passive expression is provided below, while "*-(ra)re-ta*" indicates the past tense (§2.3).

私のテスト用紙を彼に見られた。
*Watashi-no tesuto yōshi-o kare-ni mi-**rare-ta***.
My test paper was seen by him.

2.6 Causative

In English, causatives are typically expressed using "have" or "make." However, in Japanese, this can be achieved using suffixes, specifically the "*-(sa)se-ru*" form, which is annotated with CAUS (Causative).[8]

---

[5]While originally meant for contexts where a specific approver for a particular action could be anticipated, it has now changed to express humility even when the approver may not be evident (Nihongo Kijutsu Bunpo Kenkyukai, 2009b).

[6]Brackets indicate implied meaning not explicitly stated in Japanese.

[7]As in this example, the *ta*-form does not necessarily involve simply replacing "*ru*" with "*ta*" from the base form.

[8]We explain *lexical* causative verbs in §4.3.

Below is an example of the causative expression, while "*-(sa)se-ta*" indicates the past tense (§2.3).

私はその映画を彼に見させた。
*Watashi-wa sono eiga-o kare-ni mi-**sase-ta**.*
I made him watch the movie.

We also deal with the following forms: causative involving passive, and contraction of causative.

Causative and Passive   The causative expression can incorporate passivity using the "*-(sa)se-rare-ru*" form, annotated with CAUS+PASS (Causative, Passive). Below is an example of the causative and passive expression, while "*-(sa)se-rare-ta*" indicates the past tense (§2.3).

私はその映画を彼に見させられた。
*Watashi-wa sono eiga-o kare-ni mi-**sase-rare-ta**.*
I was made to watch the movie by him.
≈ He made me watch the movie.

Contraction of Causative   The contracted form "*-su/sasu*" is frequently used for causative verbs. In Regular I Verbs, similarly, the contracted form "*-sare-ru*" is commonly used for passive-causative expression (Nihongo Kijutsu Bunpo Kenkyukai, 2009a). Examples of each are presented in Appendix B.

These shortening forms, "*-su/sasu*" or "*-sare-ru*," are assigned the same labels as "*-(sa)se-ru*" (CAUS) or "*-(sa)se-rare-ru*" (CAUS+PASS). This is because they do not lead to any change in meaning, such as a decrease in respect. Below are examples of causative of the verb "*tabe-ru*" (eat-PRS).

食べる/ *tabe-ru*	食べる/ *tabe-ru*
食べさせる/ *tabe-sase-ru*	食べさす/ *tabe-sasu*
V;PRS;IPFV;**CAUS**	V;PRS;IPFV;**CAUS**

3   How to Generate Inflected Forms

The previous section outlined how we matched inflected forms with their UniMorph labels. In this section, we will walk through our process for generating all the inflected forms and how we filter out the less common forms, yielding a total of 12,687.

3.1   Seed Verb Selection Process

The selection of seed verbs (Table A in Figure 1) comprised two categories: (a) 107 basic verbs frequently encountered at the N5 (most basic) level of the Japanese Language Proficiency Test

(JLPT), and (b) 40 lexical honorifics,[9] divided into 19 respectful and 21 humble forms, as cited in Hirabayashi and Hama (1988). The number of verbs for each conjugation type and their detailed statistics are provided in Appendix C.

3.2   Generating Inflected Forms

First, we made a list of inflection patterns to be registered in J-UNIMORPH (Table B in Figure 1). Inflection patterns were carefully selected by four native speakers of Japanese (the authors), who referred to several books on Japanese grammar (Nihongo Kijutsu Bunpo Kenkyukai, 2007, 2009a,b; Hirabayashi and Hama, 1988; Takami, 2011) and a book designed for Japanese language learners (Kamiya, 2001).

Next, we used `kamiya-codec`,[10] a verb inflection tool, to generate each inflected form based on patterns derived from Kamiya (2001). This tool produces inflected forms by taking the seed verb (lemma) and the arguments for its inflections.[11] In certain cases, we modified parts of the inflected forms for additional inflection beyond what this tool provides (see Table C in Figure 1). Irregular verbs were generated manually to ensure accuracy.

Note that the definition of Japanese "word" has been controversial (Murawaki, 2019). Typically, inflected verb forms correspond to the "syntactic word" or "*bunsetsu*," a Japanese grammatical unit roughly equivalent to an English verb phrase. However, the inflected forms sometimes extend beyond this unit, especially when multiple suffixes are combined (cf. Goldman and Tsarfaty (2022)).

3.3   Filtering

To ensure the correctness and actual usage of the generated inflected forms, we used SerpAPI[12] to obtain the number of exact match 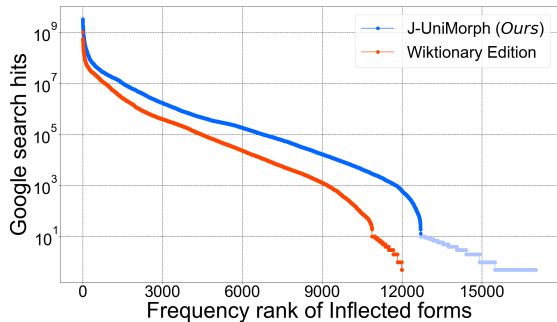hits from Google search results (Table E in Figure 1). Figure 2 shows the relationship between the frequency rank of inflected forms and their corresponding number of Google search hits, highlighting a long-tail distribution pattern. We see that the trend distinctly shifts when the number of hits reaches 10. After manually reviewing inflected forms with less than or

---

[9]Lexical honorifics are matched with the corresponding 107 basic verbs.

[10]https://github.com/fasiha/kamiya-codec

[11]One exception is the negation of "*ar-u*" (ある, be), which is expressed as "*nai*" (ない) instead of "*ar-anai*." This is implemented by `kamiya-codec`.

[12]https://serpapi.com/

Figure 2: The relationship between the frequency rank of inflected forms and their corresponding number of Google search hits, highlighting a long-tail distribution pattern, regarding J-UNIMORPH and Wiktionary Edition, respectively. Both graphs exhibit a clear trend shift when the number of hits falls to 10^1 or fewer.[15] Upon manual review by authors, for J-UNIMORPH, we concluded that these forms sound unnatural and should be discarded (indicated by the light-blue-colored plots), leaving a total of 12,687 inflected forms in J-UNIMORPH. Additionally, we found that inflected forms in Wiktionary Edition have fewer hits compared to those in J-UNIMORPH (detailed in §4.1).

equal to 10 hits, we concluded that most of these forms sound unnatural and should be discarded.[13]

We also manually removed 16 specific forms that were considered inappropriate with respect to honorifics.[14] Automating the detection and filtering of such instances will be the focus of future work.

4 Analysis of J-UNIMORPH

4.1 Comparison with Wiktionary Edition

The SIGMORPHON–UniMorph 2023 Shared Task 0 (Goldman et al., 2023) introduced a dataset focusing on Japanese Morphology, automatically extracted from Wiktionary.

Table 3 shows a comparison between the Wiktionary Edition and J-UNIMORPH in terms of the total number of inflected forms and the number of seed words. J-UNIMORPH has 12,687 inflected forms in total, which slightly exceeds the number

---

[13]We release all the generated forms with their number of Google search hits for reference.

[14]These are respectful forms of "*shinu*" (死ぬ, die) such as "*o-shini-ni-naru*" and "*shina-reru*," which sounds inappropriate and rather unnatural. A more considerate expression is "*nakunaru*" (亡くなる, pass away), which is not registered in the current version. While there are other expressions that may not be commonly used in practice, the expressions related to "die" were singled out for special attention and deletion, given the need for extra caution.

[15]To ensure visibility for forms with zero hits, we apply a smoothing technique by adding 0.5 for such cases.

Figure 3: Screenshot of J-UNIMORPH Visualizer, a tool for helping Japanese learners. Users input an inflected form and click the "Search" button to highlight corresponding UniMorph labels. If the inflected form has multiple meanings, they are displayed under the "Search Results" section, with the option to toggle between meanings. Additionally, "Related Words" section displays other inflected forms with the same label (including itself). Confidence values, ranging from 0 to 100 and based on Google search hits, assist users in determining which inflected form should be used. Higher values indicate more hits. Users also can switch between labels to investigate inflected forms with different meanings.

found in the Wiktionary Edition (12,000). We emphasize that all seed words in J-UNIMORPH are verbs, in contrast to Wiktionary Edition, where denominal verbs dominate approximately 70%. As explained in §2, inflection patterns of denominal verbs are morphologically same as those of the verb "*suru*." Table 3 also indicates that J-UNIMORPH includes a wider variety of inflection patterns and combinations, with an average of 118.6 patterns per verb, compared to the Wiktionary Edition, which averages 12.0.

Figure 2 presents the comparison of the number of Google search hits for all inflected forms listed in J-UNIMORPH and Wiktionary Edition. The graph demonstrates that J-UNIMORPH contains inflected forms that are more commonly used, as indicated by higher search hits than those in Wik-

| | Wiktionary Edition | | | J-UNIMORPH |
	Train	*Dev*	*Test*	*(Ours)*
Number of inflected forms	10,000	1,000	1,000	12,687
Number of inflected forms per word	**12.5**	**10.0**	**10.0**	**118.6**
The average of number of hits (in millions)	3.4	4.6	5.5	12
Number of seed words	800	100	100	107
Verbs	**25%**	**27%**	**30%**	**100%**
Denominal verbs (noun + "*suru*")	72%	69%	67%	0%
Accompanied by particles	3%	2%	3%	0%
Deadverbal verbs (adverb + "*suru*")	1%	2%	0%	0%

Table 3: Comparison of lemma types between Wiktionary Edition and J-UNIMORPH.

tionary Edition. The average hits by J-UNIMORPH and Wiktionary Edition are shown in Table 3.

4.2 J-UNIMORPH Visualizer

We developed the J-UNIMORPH Visualizer,[16] which takes an inflected form as the input and provides the UniMorph labels of its form (Figure 3). This makes manual analysis of J-UNIMORPH easier. Our visualizer is different from the `kamiya-codec` by accepting input with Uni-Morph labels such as Past, Negative, and Polite, instead of surface forms (*-ta*, *-nai*, *-masu*), making it more accessible to non-native users who may not be knowledgeable about surface forms and their meanings. While this tool is specifically designed for Japanese, it could be adapted to other languages with minor modifications. We hope that this visualizer can also offer a user-friendly interface for Japanese learners, enabling them to easily understand complex Japanese verb inflection patterns.

4.3 Labels and Forms Excluded from the Current Version

While J-UNIMORPH contains a total of 12,687 inflected forms, covering a variety of labels and forms as described in §2, we have excluded several forms, such as subsidiary verbs, question expressions, lexical causative verbs, and informal expressions. The primary reason for their exclusion is their simple morphological pattern or morphological equivalence to other verbs already included in J-UNIMORPH. The detailed reasons for the exclusion of these forms are provided in Appendix D.

4.4 UniMorph Limitations for Japanese

While the UniMorph schema includes a variety of morpho-semantic features, we have identified certain Japanese expressions that are not covered by the current UniMorph labels and format. In particular, due to its agglutinative nature, Japanese language includes compound suffixes consisting of multiple suffixes merging to express a new meaning beyond a simple combination of their individual semantic features (Morita and Matsuki, 1989). For example, "-*kamo-shire-nai*" ($\approx$ maybe) consists of "*kamo*" + "*shire*" + "*nai*." The full meaning emerges when these suffixes are combined, with the meaning of "*nai*" (NEG) disappearing in the process.

Importantly, the order of these suffixes matters. Below, two examples showcase the same labels (PST, PFV, and LKLY) but in a different sequence.

(a) 彼はリンゴを食べたかもしれない。
*Kare-wa ringo-o tabe-**ta-kamo-shire-nai**.*
$\approx$ He might have eaten an apple.

(b) 彼はリンゴを食べるかもしれなかった。
*Kare-wa ringo-o tabe-**ru-kamo-shire-naka-tta**.*
$\approx$ He could have been able to eat an apple.

In the example (a), the suffix "-*(t)ta*" indicates PST;PFV and "-*kamo-shire-*[*nai*|*naka*]" represents likelihood (LKLY). Although both examples contain the same set of suffixes, the meaning of each sentence differs due to the varying order of the suffixes. That is, in example (a), LKLY dominates the overall meaning more than PST+PFV, whereas in example (b), PST+PFV governs the overall meaning more than LKLY.

One approach to address this morphological complexity is to adopt a hierarchical structure for annotations, as proposed by Guriel et al. (2022), who explored complex argument marking in the Georgian language.

5 Conclusion

We introduced J-UNIMORPH, a Japanese Morphology dataset based on the UniMorph schema. J-UNIMORPH covers a wide range of verb inflection forms, including honorifics, politeness levels, and other linguistic nuances, reflecting the language's agglutinative nature. Unlike the Wiktionary Edition, which is automatically extracted from Wiktionary, J-UNIMORPH has been carefully designed by native speakers, featuring an average of 118 inflected forms per word (with a total of 12,687 instances), compared to Wiktionary Edition's 12 inflected forms per word (12,000 instances in total). J-UNIMORPH, along with its interactive visualizer, has been released to facilitate cross-linguistic research and applications, offering a more comprehensive resource than previously available.

References

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Arturo Oncevay, Juan López Bautista, Gema Celeste Silva Villegas, Lucas Torroba Hennigen, Adam Ek, David Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritván Karahóǧa, Stella Markantonatou, George Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czarnowska, Irene Nikkarinen, Aelita Salchak, Brijesh Bhatt, Christopher Straughn, Zoey Liu, Jonathan North Washington, Yuval Pinter, Duygu Ataman, Marcin Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoehr, Hossep Dolatian, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra Serova, Emily Prud'hommeaux, Maria Nepomniashchaya, Fausto Giunchiglia, Eleanor Chodroff, Mans Hulden, Miikka Silfverberg, Arya D. McCarthy, David Yarowsky, Ryan Cotterell, Reut Tsarfaty, and Ekaterina Vylomova. 2022. UniMorph 4.0: Universal Morphology. In *Proceedings of the Thirteenth Language Resources and Evaluation Confer-ence*, pages 840–855, Marseille, France. European Language Resources Association.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27, Brussels. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2017. CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 1–30, Vancouver. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The SIGMORPHON 2016 shared Task—Morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 10–22, Berlin, Germany. Association for Computational Linguistics.

Omer Goldman, Khuyagbaatar Batsuren, Salam Khalifa, Aryaman Arora, Garrett Nicolai, Reut Tsarfaty, and Ekaterina Vylomova. 2023. SIGMORPHON–UniMorph 2023 shared task 0: Typologically diverse morphological inflection. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 117–125, Toronto, Canada. Association for Computational Linguistics.

Omer Goldman and Reut Tsarfaty. 2022. Morphology Without Borders: Clause-Level Morphology. *Transactions of the Association for Computational Linguistics*, 10:1455–1472.

David Guriel, Omer Goldman, and Reut Tsarfaty. 2022. Morphological reinflection with multiple arguments: An extended annotation schema and a Georgian case study. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 196–202, Dublin, Ireland. Association for Computational Linguistics.

Yoshisuke Hirabayashi and Yumiko Hama. 1988. *Keigo (Honorific Speech)*. Aratake Publishers.

Eunkyul Jo, Kim Kyuwon, Xihan Wu, KyungTae Lim, Jungyeul Park, and Chulwoo Park. 2023. K-UniMorph: Korean Universal Morphology and its feature schema. In *Findings of the Association for*

Computational Linguistics: ACL 2023, pages 6613–6623, Toronto, Canada. Association for Computational Linguistics.

Taeko Kamiya. 2001. *The handbook of Japanese verbs*. Kodansha.

Yasuhiko Kato and Tsutomu Fukuchi. 1989. *Tense, Aspect, and Mood*. Aratake Publishers.

Jordan Kodner, Salam Khalifa, Khuyagbaatar Batsuren, Hossep Dolatian, Ryan Cotterell, Faruk Akkus, Antonios Anastasopoulos, Taras Andrushko, Aryaman Arora, Nona Atanalov, Gábor Bella, Elena Budianskaya, Yustinus Ghanggo Ate, Omer Goldman, David Guriel, Simon Guriel, Silvia Guriel-Agiashvili, Witold Kieraś, Andrew Krizhanovsky, Natalia Krizhanovsky, Igor Marchenko, Magdalena Markowska, Polina Mashkovtseva, Maria Nepomniashchaya, Daria Rodionova, Karina Scheifer, Alexandra Sorova, Anastasia Yemelina, Jeremiah Young, and Ekaterina Vylomova. 2022. SIGMORPHON–UniMorph 2022 shared task 0: Generalization and typologically diverse morphological inflection. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 176–203, Seattle, Washington. Association for Computational Linguistics.

Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2020. UniMorph 3.0: Universal Morphology. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3922–3931, Marseille, France. European Language Resources Association.

Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence, Italy. Association for Computational Linguistics.

Yoshiyuki Morita and Masae Matsuki. 1989. *Nihongo Hyogen Bunkei (Structures of Japanese Expressions)*. ALC PRESS.

Yugo Murawaki. 2019. On the Definition of Japanese Word. *ArXiv*, abs/1906.09719.

Nihongo Kijutsu Bunpo Kenkyukai. 2007. *Gendai Nihongo Bunpo 3 (Contemporary Japanese Grammar 3)*. Kurosio Publishers. (In Japanese).

Nihongo Kijutsu Bunpo Kenkyukai. 2009a. *Gendai Nihongo Bunpo 2 (Contemporary Japanese Grammar 2)*. Kurosio Publishers. (In Japanese).

Nihongo Kijutsu Bunpo Kenkyukai. 2009b. *Gendai Nihongo Bunpo 7 (Contemporary Japanese Grammar 7)*. Kurosio Publishers. (In Japanese).

Tiago Pimentel, Maria Ryskina, Sabrina J. Mielke, Shijie Wu, Eleanor Chodroff, Brian Leonard, Garrett Nicolai, Yustinus Ghanggo Ate, Salam Khalifa, Nizar Habash, Charbel El-Khaissi, Omer Goldman, Michael Gasser, William Lane, Matt Coler, Arturo Oncevay, Jaime Rafael Montoya Samame, Gema Celeste Silva Villegas, Adam Ek, Jean-Philippe Bernardy, Andrey Shcherbakov, Aziyana Bayyr-ool, Karina Sheifer, Sofya Ganieva, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Andrew Krizhanovsky, Natalia Krizhanovsky, Clara Vania, Sardana Ivanova, Aelita Salchak, Christopher Straughn, Zoey Liu, Jonathan North Washington, Duygu Ataman, Witold Kieraś, Marcin Woliński, Totok Suhardijanto, Niklas Stoehr, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Richard J. Hatcher, Emily Prud'hommeaux, Ritesh Kumar, Mans Hulden, Botond Barta, Dorina Lakatos, Gábor Szolnok, Judit Ács, Mohit Raj, David Yarowsky, Ryan Cotterell, Ben Ambridge, and Ekaterina Vylomova. 2021. SIGMORPHON 2021 shared task on morphological reinflection: Generalization across languages. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–259, Online. Association for Computational Linguistics.

John Sylak-Glassman. 2016. The composition and use of the universal morphological feature schema (unimorph schema).

Ken-ichi Takami. 2011. *Ukemi to Shieki (Passive and Causative)*. Kaitakusha.

Ekaterina Vylomova, Jennifer White, Elizabeth Salesky, Sabrina J. Mielke, Shijie Wu, Edoardo Maria Ponti, Rowan Hall Maudslay, Ran Zmigrod, Josef Valvoda, Svetlana Toldova, Francis Tyers, Elena Klyachko, Ilya Yegorov, Natalia Krizhanovsky, Paula Czarnowska, Irene Nikkarinen, Andrew Krizhanovsky, Tiago Pimentel, Lucas Torroba Hennigen, Christo Kirov, Garrett Nicolai, Adina Williams, Antonios Anastasopoulos, Hilaria Cruz, Eleanor Chodroff, Ryan Cotterell, Miikka Silfverberg, and Mans Hulden. 2020. SIGMORPHON 2020 shared task 0: Typologically diverse morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–39, Online. Association for Computational Linguistics.

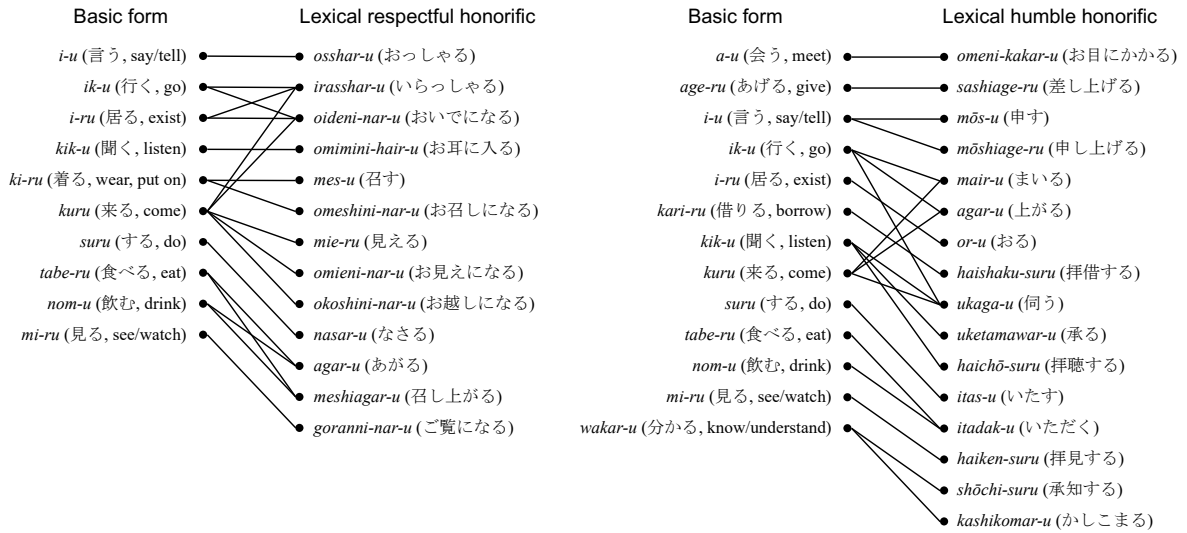## A Correspondence between the basic form and the lexical honorifics

Figure 4: Correspondence between the basic forms and the lexical honorifics adopted in J-UNIMORPH.

B Examples of Contraction form of Causative

Conj. type	Base	Ordinary	Contraction
Reg. I	書く	書かせる	書かす
	kak-u	*kak-ase-ru*	*kak-as-u*
Reg. II	見る	見させる	見さす
	mi-ru	*mi-sase-ru*	*mi-sas-u*
Irreg.	来る	来させる	来さす
	ku-ru	*ko-sase-ru*	*ko-sas-u*
Irreg.	する	させる	さす
	su-ru	*s-ase-ru*	*sas-u*

Table 4: Examples of Causative contraction forms. We also handle these contraction forms.

Conj. type	Base	Ordinary	Contraction
Reg. I	書く	書かせられる	書かされる
	kak-u	*kak-ase-rare-ru*	*kak-as-are-ru*
Reg. II	見る	見させられる	*見さされる
	mi-ru	*mi-sase-rare-ru*	*mi-sas-are-ru*
Irreg.	来る	来させられる	*来さされる
	ku-ru	*ko-sase-rare-ru*	*ko-sas-are-ru*
Irreg.	する	させられる	*さされる
	su-ru	*s-ase-rare-ru*	*sas-are-ru*

Table 5: Examples of Passive-Causative contraction forms. We do not handle incorrect usages, which have the asterisk (*).

C Statistics of generated inflected forms in J-UNIMORPH

Politeness Type	Conjugation Type	Verbs	Generated inflected forms
Basic	Regular I	76	126
	Regular II	29	118
	"*kuru*" (Irregular)	1	100
	"*suru*" (Irregular)	1	102
Lexical respectful honorifics	Regular I	18	103
	Regular II	1	94
Lexical humble honorifics	Regular I	15	92
	Regular II	2	84
	"*-suru*" (Irregular)	4	84

Table 6: The number of verbs and generated inflected forms per verb for each conjugation type. The numbers represent the counts prior to excluding infrequent inflected forms.

D Inflection/derivation affixes not included in J-UɴɪMᴏʀᴘʜ

We provide several details on the excluded forms in J-UɴɪMᴏʀᴘʜ, with the detailed list available in Table 8.

Subsidiary Verbs   In Japanese, a small group of verbs, referred to as subsidiary verbs, are characterized by their grammaticalized functions after the *te*-form. Subsidiary verbs contribute additional meanings to the verbs they attach to. For example, the verb "*iru*," conveying "be" independently, transforms into "be running" or "have run" in the context of "*hashi-tte-iru*." Similarly, the verb "*miru*," meaning "look" or "watch" on its own, takes on a different meaning, such as "try running," when attached to the verb "*hashi-ru*" (run) like "*hashi-tte-miru*." We generally excluded subsidiary verbs from J-UɴɪMᴏʀᴘʜ due to their morphological equivalence to the subsidiary verbs that are already incorporated into J-UɴɪMᴏʀᴘʜ as seed verbs. Furthermore, one subsidiary verb can precede another subsidiary verb, to express a wide range of possible combinations, such as "*hashi-tte-mi-te-iru*." We set aside these patterns for future research.

Question Expressions   The interrogative (INT) suffix "*ka*" forms questions,[17] easily added to create inflected forms. However, its use with other suffixes can alter meanings. For example, "*tabe-masen*" (eat-PRS;POL;NEG), meaning "(I) don't eat," becomes "Shall (we) eat?" when "*ka*" is added, as in "*tabe-masen-ka*?" (eat-INT;INTEN;POL), dropping the negation. Matching these combined forms with their meanings is complex, and we reserve this for future research.

Lexical causative verbs   In addition to verbs that marked CAUS (Causative) by attaching "*-se-ru/sase-ru*" (§2.6), some verbs have the corresponding transitive forms that inherently carry both the causation process and the resulting event (Takami, 2011). Below, example (a) shows the base form "*ne-ru*" (寝る, sleep) with the causative inflection suffix, whereas example (b) uses lexical causative verb "*nekas-u/nekas-e-ru*" (寝かす/寝かせる, make someone sleep) to express causative feature. We did not include lexical causative verbs in J-UɴɪMᴏʀᴘʜ because they are not expressed through inflection.

(a) お母さんは子供を寝させた。
*Okāsan-wa, kodomo-o ne-**sase-ta***. ("*-sase-ru*" form)
The mother put the child to sleep.

(b) お母さんは子供を寝かした/寝かせた。
*Okāsan-wa, kodomo-o **nekash-i-ta/nekas-e-ta***. (lexical causative verb)
The mother put the child to sleep.

Controversial Informal Language Form   Several colloquial expressions are controversial and seen as incorrect in Japanese.[18] Table 7 shows examples of omitting "*ra*," omitting "*i*," and inserting "*sa*." Although these expressions are widely used in spoken language, they are not currently used in newspapers and formal writings, and are still considered incorrect in standard language. Therefore, we have excluded them from the current version of J-UɴɪMᴏʀᴘʜ.

Special usage of *ru-* and *ta*-form   The *ru-* and *ta*-form, which were mentioned in §2.3, have various meanings by being accompanied by peripheral words such as adverbs and interjections. The examples about special usage of the *ru*-form are property: 日本人は米を食べる。 (Japanese people eat rice.), and command: さっさと歩く！ (Walk quickly!). The examples about special usage of the *ta*-form are discovery: [鍵を探していて] あっ、ここにあった。 (Oh, here's the key.), and recall: あっ、今日は会議だった。 (Oh, I have a meeting today.) (Nihongo Kijutsu Bunpo Kenkyukai, 2007). Since the meaning of these cases relies on peripheral words, not on the inflected form itself, we exclude these instances from the J-UɴɪMᴏʀᴘʜ.

---

[17]In conversational contexts, raising the intonation at a sentence's end can indicate a question without a specific marker.

[18]https://www.bunka.go.jp/kokugo_nihongo/sisaku/joho/joho/kakuki/20/tosin03/09.html

Category	Formal Form	Informal Form	Rough translation
Omitting *ra*	*tabe-**ra**reru* /食べられる	*tabe-reru* /食べれる	can eat
Omitting *i*	*tabe-te-iru* /食べている	*tabe-te-ru* /食べてる	be eating, have eaten
Inserting *sa*	*kawa-sete-itadaku* /買わせていただく	*kawa-**sa**-sete-itadaku* /買わさせていただく	have the honor of buying

Table 7: Examples of Informal Forms

Reason	Affixes or example Inflected forms	Romanized and Rough translation
Subsidiary verbs (補助動詞)	〜ている	*-te-iru* (be doing, have done)
	〜てみる	*-te-miru* (try doing)
	〜ておく	*-te-oku* (do in advance)
	〜ておこう	*-te-okō* (let's do in advance)
	〜てあげる	*-te-ageru* (do something for the benefit of someone)
	〜てもらう	*-te-morau* (get someone to do something)
	〜てくれる	*-te-kureru* (someone do something for me/us)
	〜てある	*-te-aru* (has been done)
	〜てしまう	*-te-shimau* (end up doing)
	〜ていく	*-te-iku* (keep on doing)
	〜つつある	*-tsutsu-aru* (be about to do)
	〜てほしい	*-te-hoshii* (want someone to do)
Compound suffixes (複合辞)	〜かもしれない	*-kamo-shire-nai* (may)
	〜てはいけない	*-tewa-ike-nai* (must not do)
	〜てはならない	*-tewa-nara-nai* (must not do)
	〜たがっている	*-tagatte-iru* (wants to do)
	〜なければならない	*-nakereba-naranai* (have to do)
	〜に違いない	*-ni-chigai-nai* (must be doing)
Non verbs	〜てもいい	*-te-mo-ii* (permissive)
	〜たら，〜ば	*-tara*, *-ba* (if)
	〜たり	*-tari* (do and ...)
	〜べきだ，〜べし	*-beki-da*, *-beshi* (should do)
	〜つもりだ	*-tsumori-da* (intend to do)
	〜はずだ	*-hazu-da* (be supposed to do)
	〜らしい	*-rashii* (It seems like ...)
	〜べからず	*-bekara-zu* (should not do)
	「笑い」「話」など	Treat as nouns, such as *warai* (laughter), *hanashi* (talk/conversation)
	〜に〜（「買いに行く」など）	*-ni-* (adverbial usage)
	〜ながら	*-nagara* (while doing)
	〜そうだ	*-sōda* (It seems like ...)
	〜物，〜方	*-mono*, *-kata* (Nominative usage)
	〜始める，〜終わる	*-hajimeru*, *-owaru* (begin -ing, finish -ing)
Noun/Adverb + light verb	〜する	*-suru* (light verb)
Lexical causative verbs	寝かせる，立てる	*nekaseru*, *tateru*
Omitting *ra* (ら抜き言葉)	〜れる	*-reru*
Omitting *i* (い抜き言葉	〜てる	*-teru*
Inserting *sa* (さ入れ言葉)	〜させて〜	*-sase-te-*
Interrogative suffix	〜か？	*-ka?*
	〜ましょうか？，〜ませんか？	*-mashōka?*, *masen-ka?*
Another respectful expressions	お〜くださる	*o—kudasaru*
	お〜なさる	*o—nasaru*
Another humble expressions	お〜いたす	*o—itasu*
	お〜いたします	*o—itashi-masu*
Others	〜れる/られる	*-(ra)reru* (spontaneous)
	〜よう	*-yō* (speculation)

Table 8: List of inflection/derivation affixes not included in the current version of J-UNIMORPH.