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

Kosuke Matsuzaki [1], Masaya Taniguchi [2], Kentaro Inui [3,1,2], Keisuke Sakaguchi [1,2]

([1] Tohoku University, [2] RIKEN, [3] MBZUAI)

matsuzaki.kosuke.r7@dc.tohoku.ac.jp
@Matsuzaki_NLP
https://matsukosuke.github.io/

SIGMORPHON2024@Mexico City, June 20th

TOHOKU NLP GROUP 東北大学 自然言語処理 研究グループ

Overview

- **Introduction**
  - Prior Japanese dataset in UniMorph
- **J-UniMorph: Japanese Adaptation**
  - Creation process of J-UniMorph
  - Comparison with the prior dataset
- **Application: J-UniMorph Visualizer**
  - Aiming to support Japanese learners in semantic understanding and morphological usage

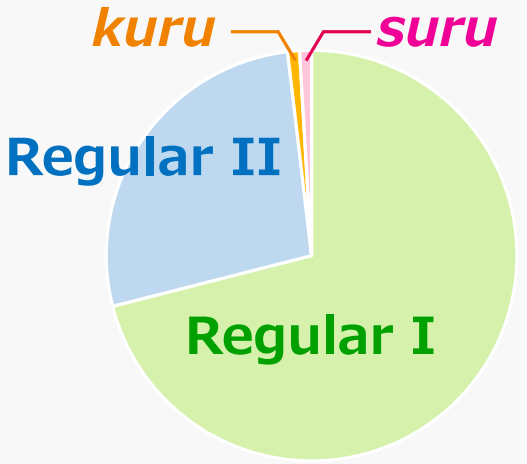
lemma	inflected form	label
eat	ate	V;PST

Illustration of UniMorph

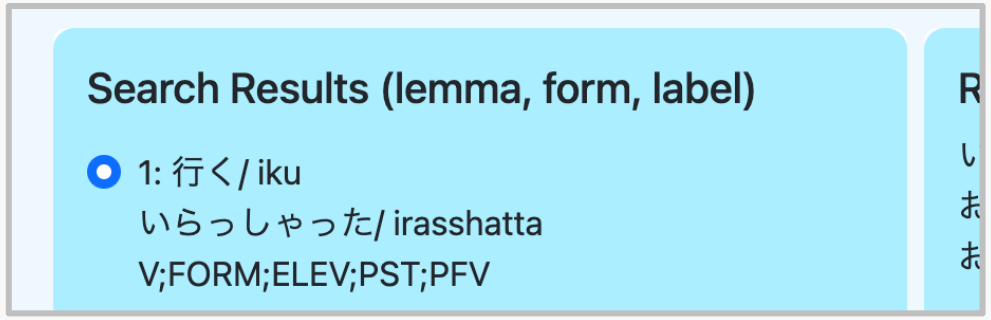
unimorph/jpn

Contributions to UniMorph [Sylak-Glassman'16]

- **Increased Japanese Inflection Patterns and Combinations**
  - The number of inflection patterns becomes about 10x larger than the prior dataset
- **Included High-Frequency Forms**
  - Average Google Search Hits became higher
- **Improved Data Accessibility**
  - J-UniMorph Visualizer made it easier to explore and analyze



J-UniMorph



Prior Japanese Dataset in UniMorph

- Released for the SIGMORPHON–UniMorph 2023 Shared Task 0
  - Automatically extracted from Wiktionary [Goldman+'23]
  - We call this **Wiktionary Edition**

- Basic statistics of the Wiktionary Edition:
  - 1. About 70% of verbs have the **same inflection pattern** †
    - These inflections can be replaced by one seed verb (lemma)
  - 2. There are an **average of 12 entries** for each verb
    - Japanese verbs have a more diverse range of inflections

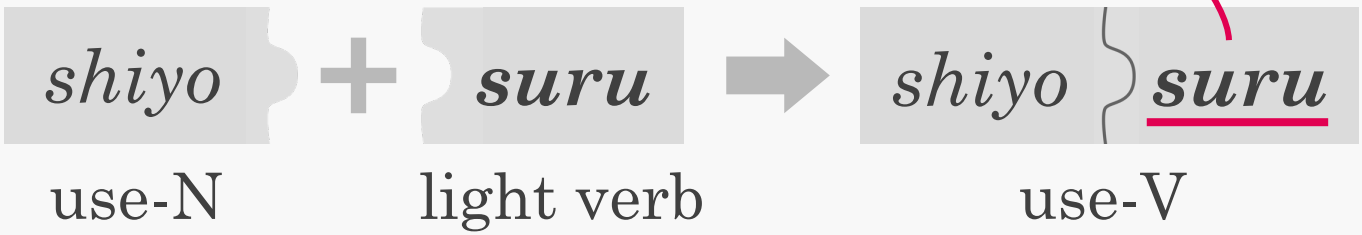
†Details on the next slide

Derived Verbs vs. Common Verbs

- **Regular I verbs**
  - e.g., *kak-u* (write-V), *hashir-u* (run-V)
- **Regular II verbs**
  - e.g., *tabe-ru* (eat-V), *mi-ru* (see-V)
- **Irregular verbs**
  - Only *kuru* (come-V), ***suru*** (do-V)

Derived verbs

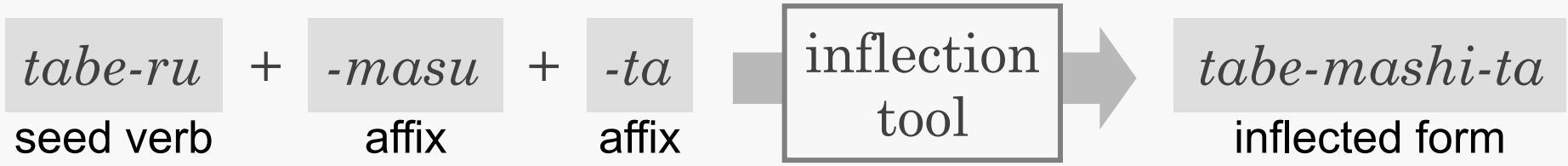
- e.g.,



Creation Process of J-UniMorph

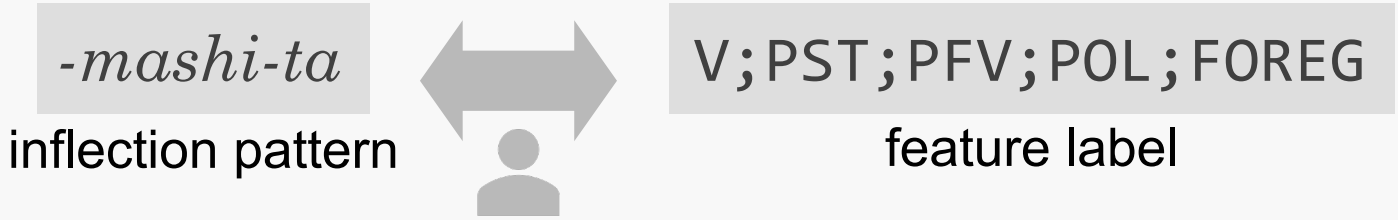
1. Inflection Generation

- Generating the forms with verb inflection tool (semi-automatically)



2. Label Annotation Mapping

- Annotating UniMorph labels to each inflection pattern

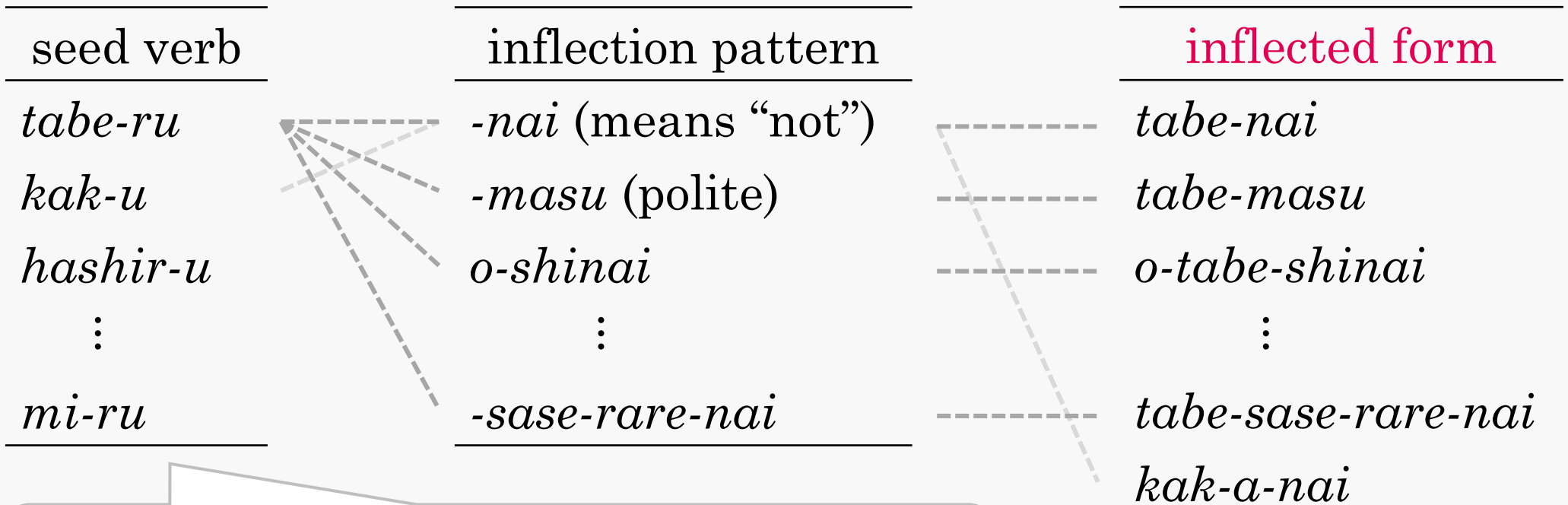


3. Filtering

- Remove incorrect or infrequent forms based on frequency

Dataset creation Step 1: Inflection Generation

- Generate inflected forms with the rule-based verb inflection tool



Dataset creation Step 2: Label Mapping

- Annotate the UniMorph labels to each inflection pattern (affixes)
  - References:
    - Definitions of UniMorph feature schema [Sylak-Glassman'16]
    - Books on Japanese grammar [Nitta'07, '09a, '09b; Hirabayashi+'88; Kato+'89; Kamiya'01; Takami'11]

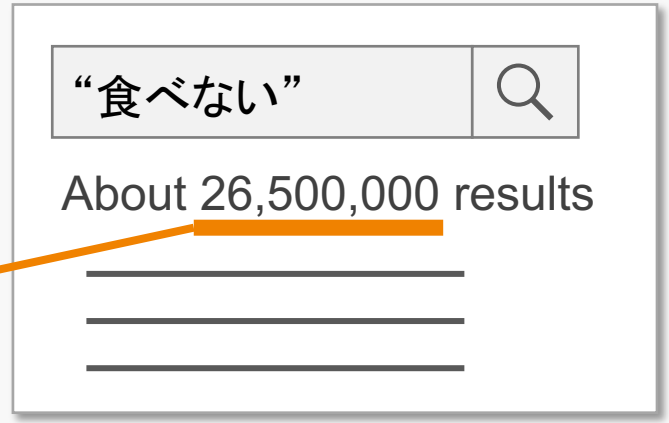
inflection pattern	UniMorph feature label
-nai	V;PRS;IPFV;NEG
-masu	V;PRS;IPFV;POL;FOREG
o-shinai	V;PRS;IPFV;FORM;HUMB;NEG
⋮	⋮
-sase-rare-nai	V;PRS;IPFV;CAUS;PASS;NEG

Dataset creation Step 3: Filtering

- Remove unnatural inflected forms
  - Using the number of hits from Google's exact match search
  - Remove forms with fewer than 10 hits

Inflected form	Romanization	Hits
食べない	*tabe-nai*	26,500,000
食べます	*tabe-masu*	22,600,000
お食べしない	*o-tabe-shinai*	6
⋮	⋮	⋮
見させられない	*mi-sase-rare-nai*	137,000

← remove (≤ 10)



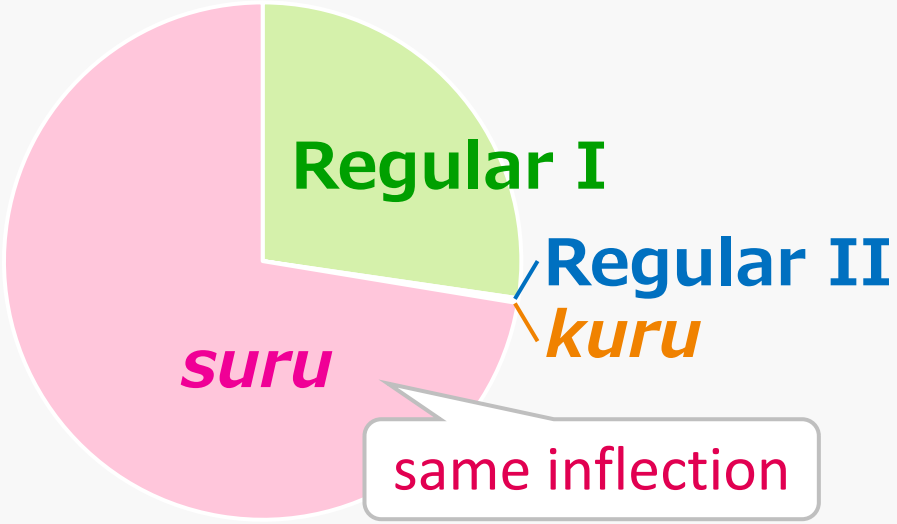
Comparison (1/3): Average Entries per Word

- J-UniMorph has various inflection patterns and combinations
  - About 10x larger than the Wiktionary Edition despite similar entries

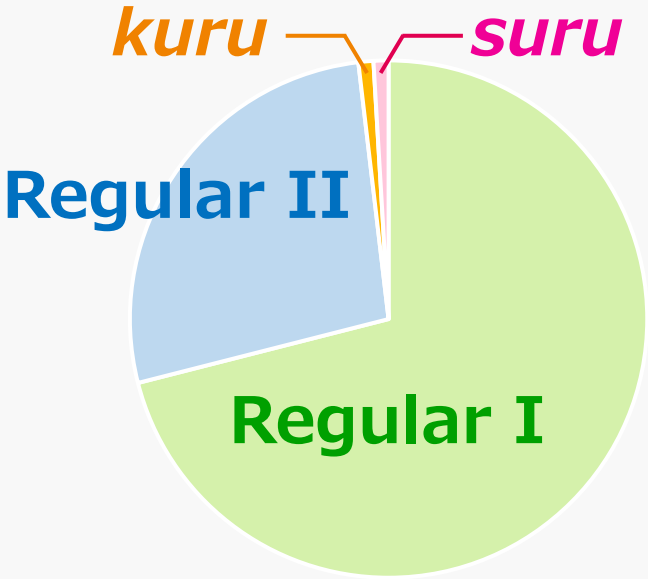
		Ours
	Wiktionary Edition	J-UniMorph
Total number of entries	12,000	12,687
Number of entries **per word**	**12.0**	**118.6**

Comparison (2/3): Verb Statistics

- Wiktionary Edition:
  - About 70% are the verbs which end with the light verb “*suru*”
  - These inflections can be replaced by the single verb “*suru*”



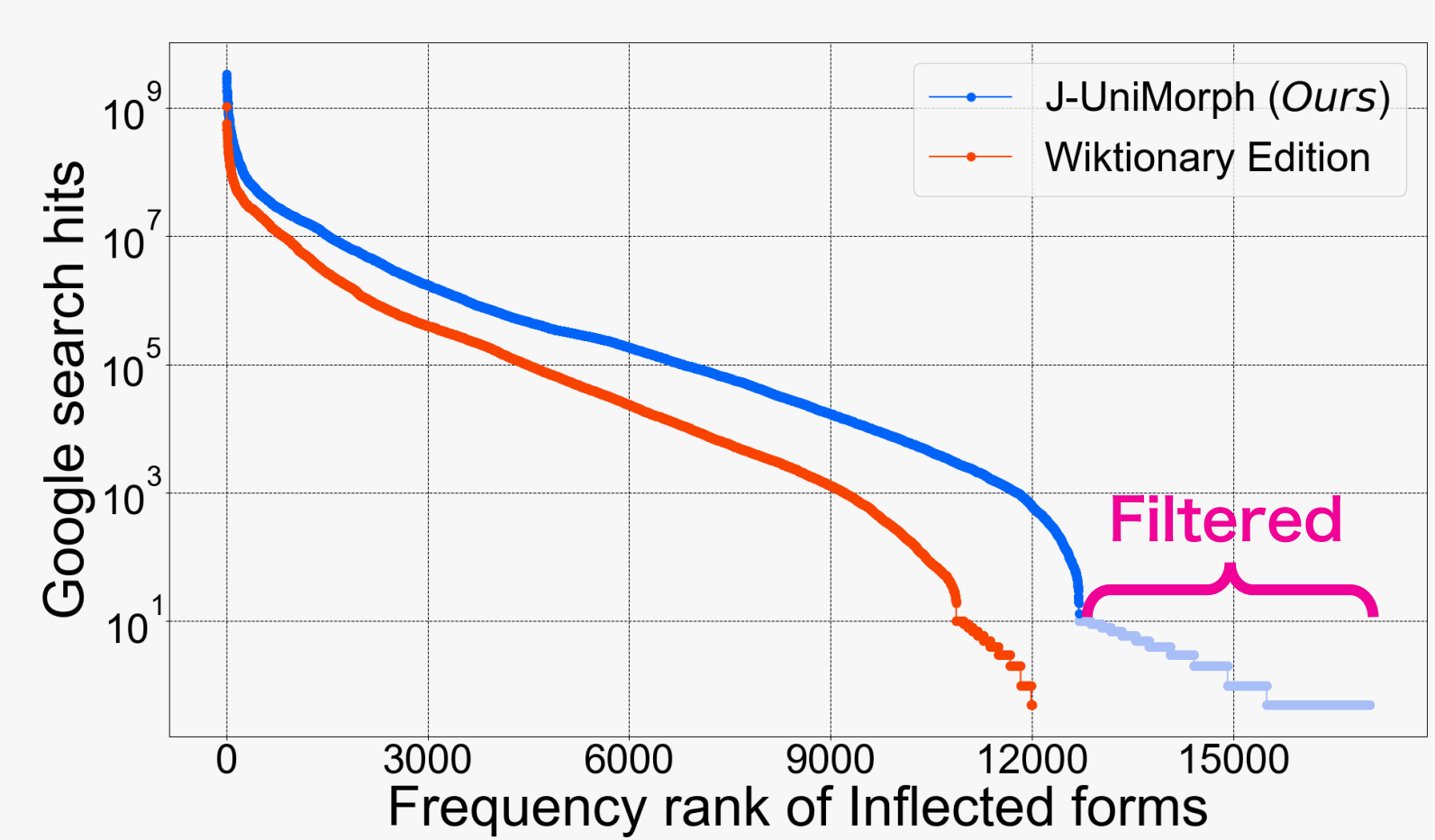
Wiktionary Edition
(1,000 verbs / 12,000 entries)



J-UniMorph
(107 verbs / 12,687 entries)

Comparison (3/3): Hits of each Inflected Form

- The entries in J-UniMorph have higher frequency forms



Application: J-UniMorph Visualizer

- Creating J-UniMorph
  - Dataset creation process
  - Comparison with the Wiktionary Edition

- **J-UniMorph Visualizer**
  - Aims to support Japanese learners in semantic understanding and morphological usage

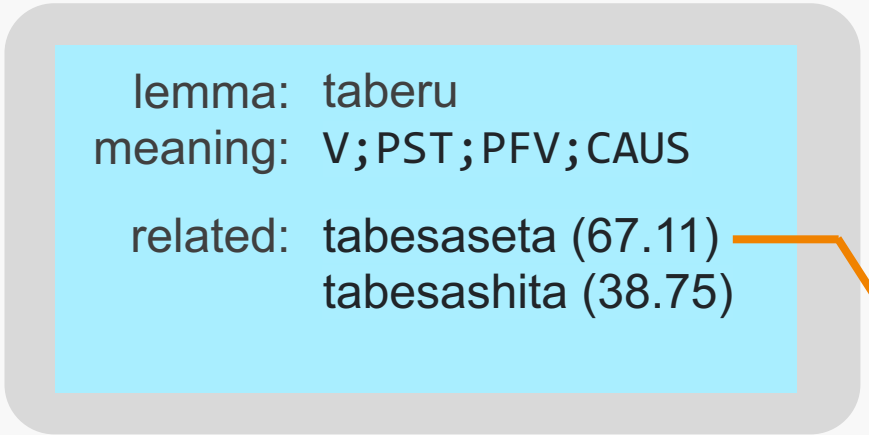
↑Visualizer

Visualizer (1/3): Semantic Understanding

You can analyze the morphological meaning of inflected forms

Input 1: an inflected form

Output 1: lemma, morphological meaning, and related words



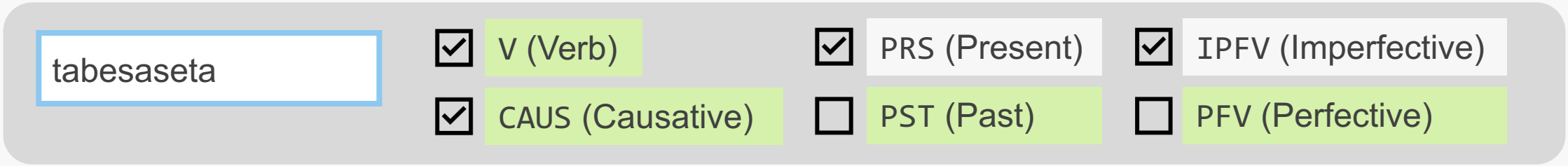
$$\text{FrequencyIndex} = 100 \times \frac{\log_{10}(\text{hits} + 1)}{\log_{10}(\text{max\_hits} + 1)}$$

“lemma” and “meaning” can be displayed in multiple instances

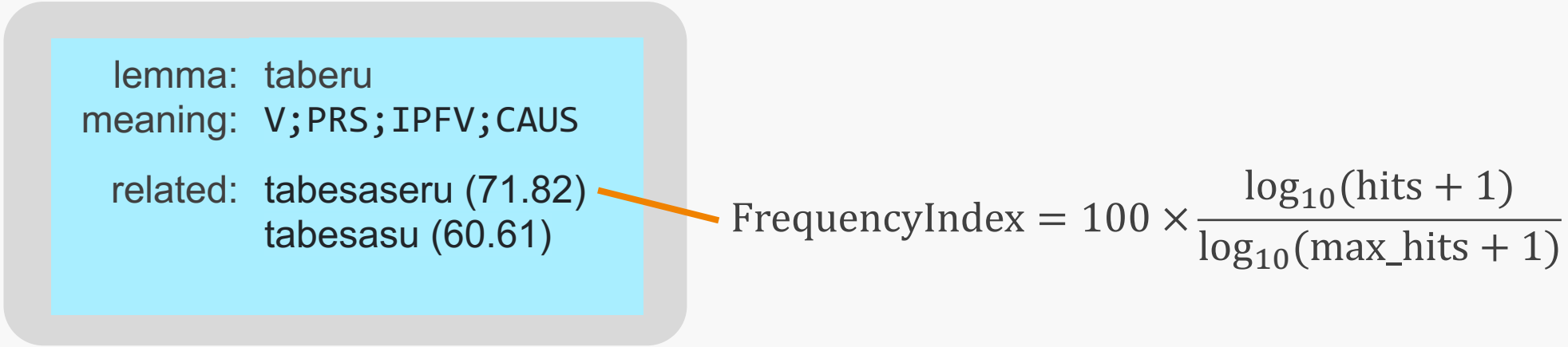
Visualizer (2/3): Morphological Usage

You can display other inflected forms with different labels

Input 2: switch the checkboxes to display other forms



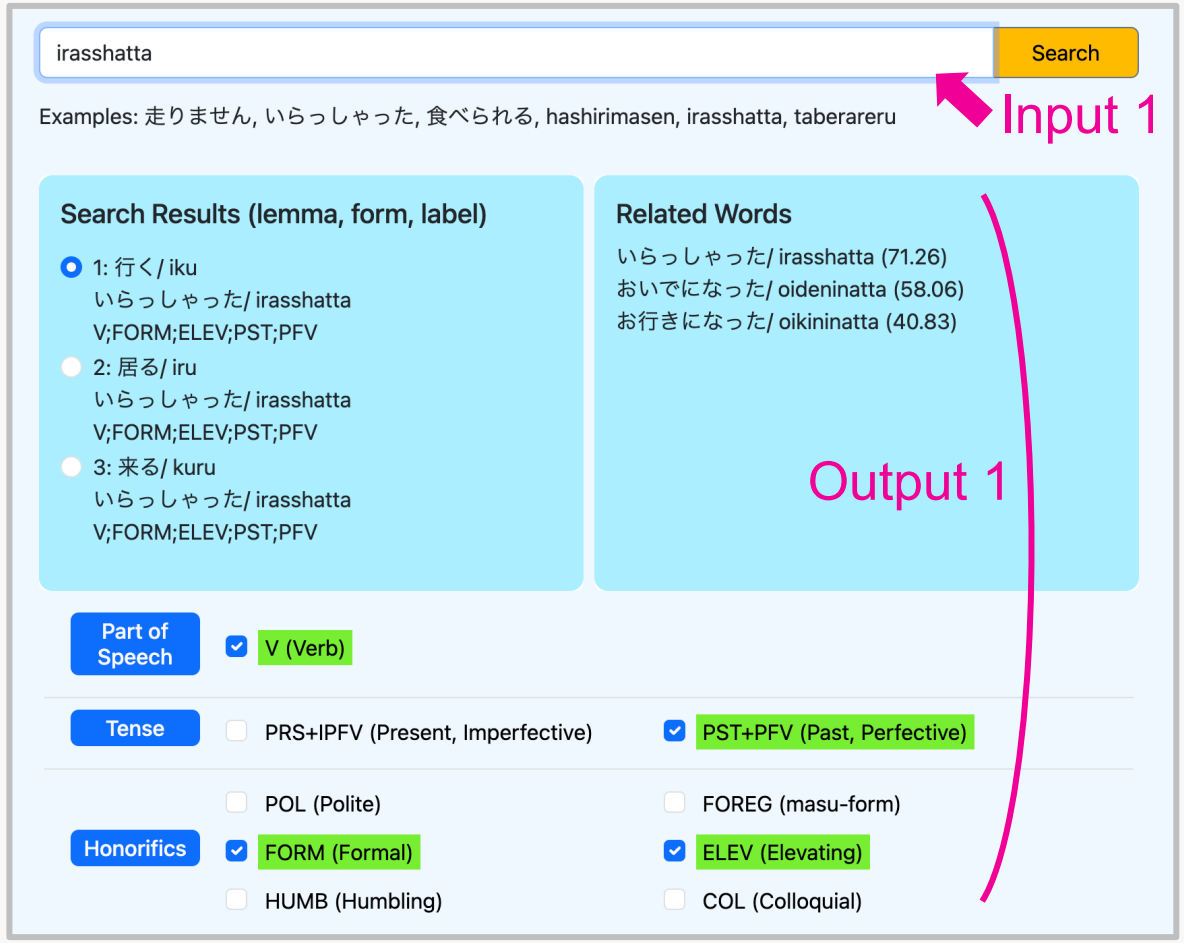
Output 2: words which have new meanings you clicked



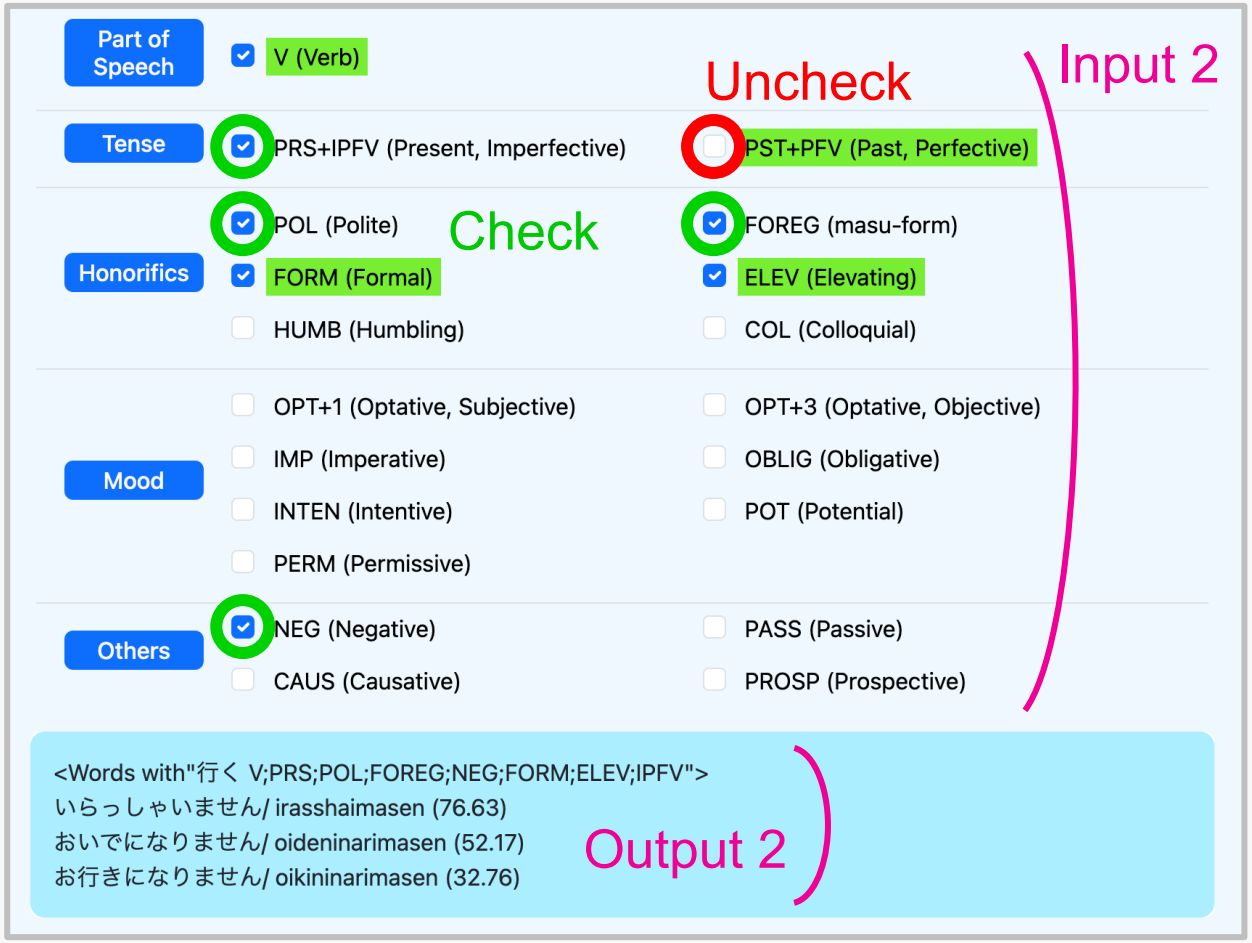
$$\text{FrequencyIndex} = 100 \times \frac{\log_{10}(\text{hits} + 1)}{\log_{10}(\text{max\_hits} + 1)}$$

Visualizer (3/3): Screenshots

1: See lemma and meaning



2: See other inflected forms



Extended dataset over UniMorph Schema

- We created the additional dataset for J-UniMorph Visualizer
  - https://github.com/cl-tohoku/J-UniMorph
- It includes (without Filtering):
  - Lemma (Kanji & **Hiragana** & **Romanized form**)
  - Inflected form (Kanji & **Hiragana** & **Romanized form**)
  - UniMorph Feature Labels
  - **Google Search Hits**

Lemma	Inflected form	Feature label	Hits
書く/ かく/ kaku	書かない/ かかない/ kakanai	V;PRS;IPFV;NEG	16000000

Summary

- We introduced **J-UniMorph**, a Japanese morphology dataset
  - Based on the UniMorph schema
  - Total: 12,687 instances
  - Covers a wide range of verb inflection forms (average: 118 / word)
    - Compared to Wiktionary Edition's 12 / word
- We developed its interactive **Visualizer** →
- Future work:
  - Adding unaddressed inflection patterns
  - Expanding coverage to nouns and adjectives

References

[Sylak-Glassman'16] John Sylak-Glassman. The composition and use of the universal morphological feature schema (unimorph schema). 2016.

[Goldman+'23] 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.

[Nitta'07] Nihongo Kijutsu Bunpo Kenkyukai. 2007. *Gendai Nihongo Bunpo 3*. Kurosio Publishers.

[Nitta'09a] Nihongo Kijutsu Bunpo Kenkyukai. 2009. *Gendai Nihongo Bunpo 2*. Kurosio Publishers.

[Nitta'09b] Nihongo Kijutsu Bunpo Kenkyukai. 2009. *Gendai Nihongo Bunpo 7*. Kurosio Publishers.

[Hirabayashi+'88] Yoshisuke Hirabayashi and Yumiko Hama. 1988. *Keigo*. Aratake Publishers.

[Kato+'89] Yasuhiko Kato and Tsutomu Fukuchi. 1989. *Tense, Aspect, and Mood*. Aratake Publishers.

[Kamiya'01] Taeko Kamiya. 2001. *The handbook of Japanese verbs*. Kodansha.

[Takami'11] Ken-ichi Takami. 2011. *Ukemi to Shieki*. Kaitakusha.

Acknowledgements

- This work was supported by RIKEN Special Postdoctoral Researchers Program and JSPS KAKENHI Grant Numbers JP21K21343, JP22H00524.
- While conducting this research, we received valuable comments from members of the Tohoku NLP Group at Tohoku University, Japan.
  We deeply appreciate their support and insightful contributions.

Thank you for listening!

- We introduced **J-UniMorph**, a Japanese morphology dataset
  - Based on the UniMorph schema
  - Total: 12,687 instances
  - Covers a wide range of verb inflection forms (average: 118 / word)
    - Compared to Wiktionary Edition's 12 / word
- We developed its interactive **Visualizer** →
- Future work:
  - Adding unaddressed inflection patterns
  - Expanding coverage to nouns and adjectives

Appendix

GitHub Repository for J-UniMorph

J-UniMorph

https://github.com/cl-tohoku/J-UniMorph

`jpn` repository in UniMorph project

https://github.com/unimorph/jpn

Introduction: What is UniMorph?

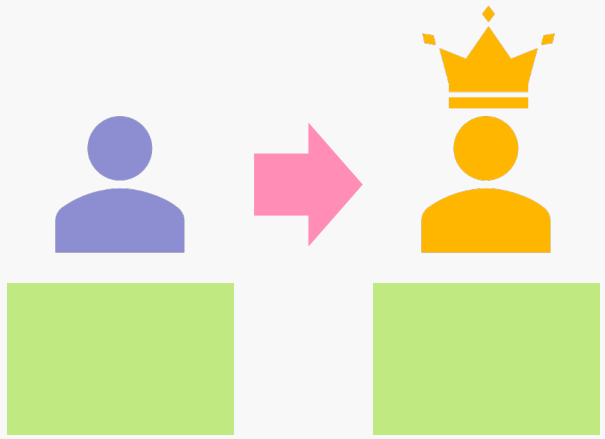
- Universal Morphology (UniMorph) [Sylak-Glassman'16]
  - Collaborative project
  - Covers over **170** languages globally

- UniMorph Feature Schema [Sylak-Glassman'16]
  - Provides standardized morphological features
    - Over 212 feature labels across 23 dimensions of meaning
  - Express with **triplets**: lemma, inflected form, and feature label
    - e.g.

lemma	inflected form	feature label
tabe-ru/食べる	*tabe-ta*/食べた	V;PST;PFV

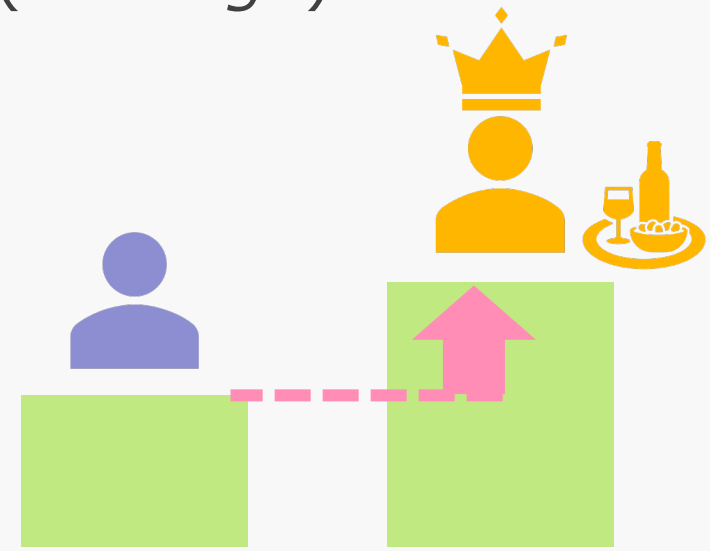
Another Feature of J-UniMorph: Honorifics

Polite form
(*Teineigo*)



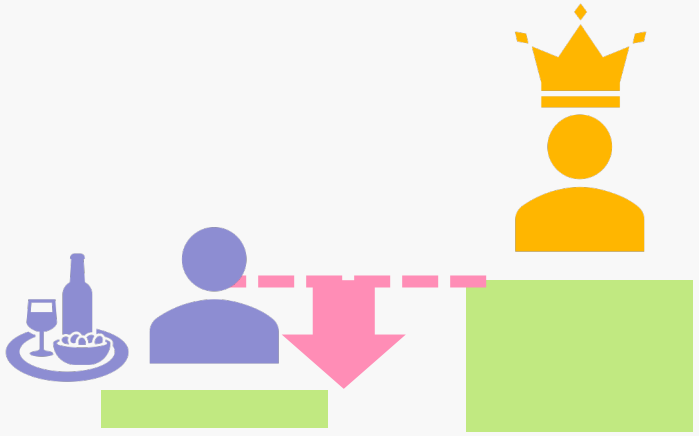
tabe-ru	eat-V
tabe-masu	POL

Respectful form
(*Sonkeigo*)



meshiaga-ru	ELEV
meshiaga-ri-masu	ELEV;POL

Humble form
(*Kenjogo*)



itada-ku	HUMB
itada-ki-masu	HUMB;POL

Honorifics do not correspond one-to-one

- We treated both simple inflections and lexical honorifics
  - Respectful forms are labeled `FORM;ELEV` (Formal, Referent Elevating)

lemma	inflected form	feature label
tabe-ru (eat)	*tabe-ta*	`V;PST;PFV`
tabe-ru (eat)	*o-tabe-ni-na-tta*	`V;FORM;ELEV;PST;PFV`
tabe-ru (eat)	*meshiaga-tta*	`V;FORM;ELEV;PST;PFV`
tabe-ru (eat)	*aga-tta*	`V;FORM;ELEV;PST;PFV`
no-mu (drink)	*non-da*	`V;PST;PFV`
no-mu (drink)	*meshiaga-tta*	`V;FORM;ELEV;PST;PFV`
no-mu (drink)	*aga-tta*	`V;FORM;ELEV;PST;PFV`

Why did we drop the Derived Verbs?

	Seed Verb (Lemma)	Negative Form
• Regular I	◦ *kak-u* (write-V)	→ *kak-a-nai*
	◦ *okur-u* (send-V)	→ *okur-a-nai*
• Regular II	◦ *tabe-ru* (eat-V)	→ *tabe-nai*
	◦ *mi-ru* (look-V)	→ *mi-nai*
• Irregular	◦ *kuru* (come-V)	→ *konai*
	◦ ***suru*** (do-V)	→ ***shinai***
• Derived Verbs	◦ *shiyo-**suru*** (use-N + do)	→ *shiyo-**shinai***
	◦ *benkyo-**suru*** (study-N + do)	→ *benkyo-**shinai***

The Feature of Japanese Morphology

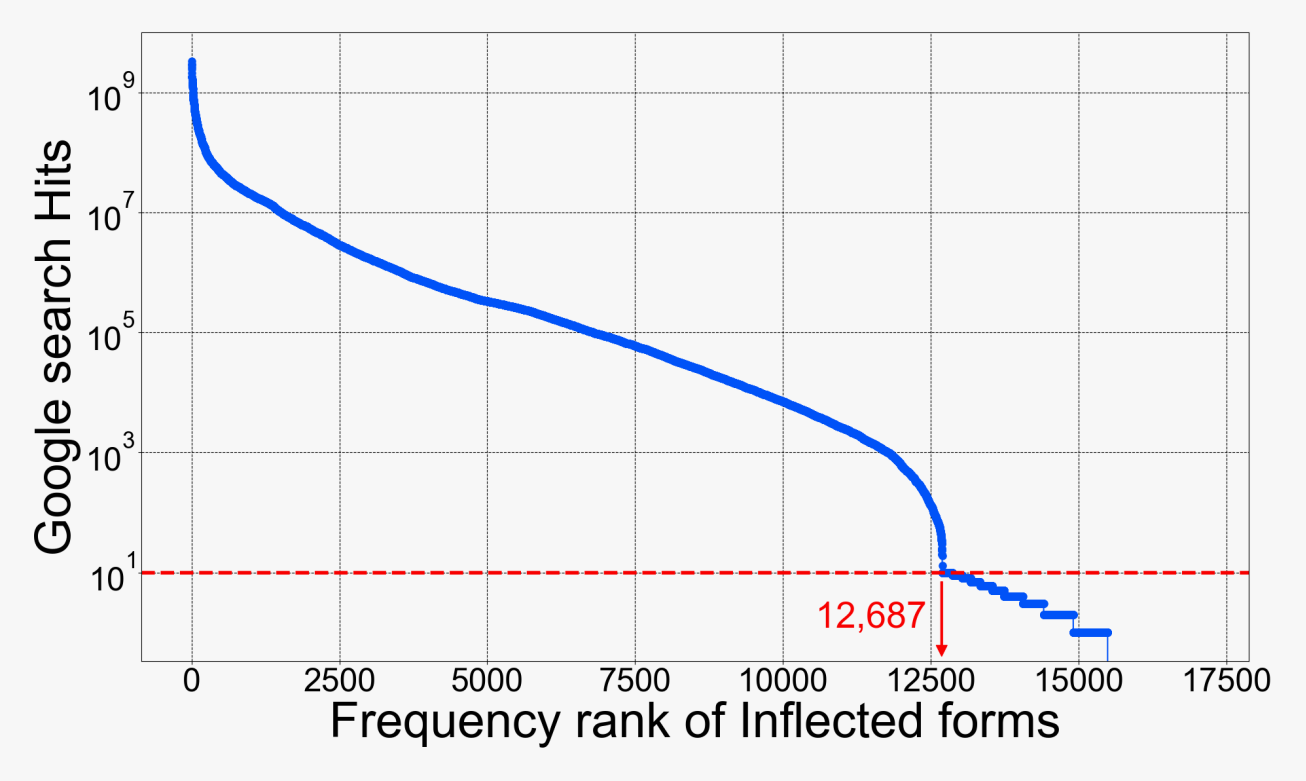
- Morphemes do not correspond one-to-one with meanings

tabe-nai-desu … (I) don't eat. (Colloquial)

lemma	inflected form	feature label
tabe-ru	*tabe-nai-desu*	V;PRS;IPFV;NEG;POL;COL

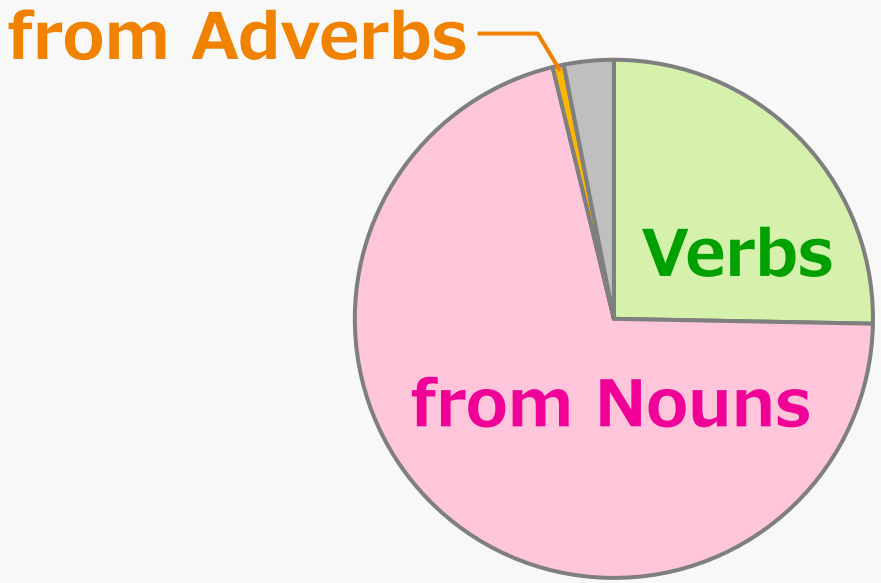
In the Filtering step: Why 10 hits?

- The trend distinctly shifts when the number of hits reaches 10
- Most of these (≤ 10) forms sound unnatural



Comparison (2/3): Verb Analysis about Derivation

- Wiktionary Edition:
  - About 70% are derived from nouns or adverbs
  - These inflection patterns are morphologically same as “*suru*”



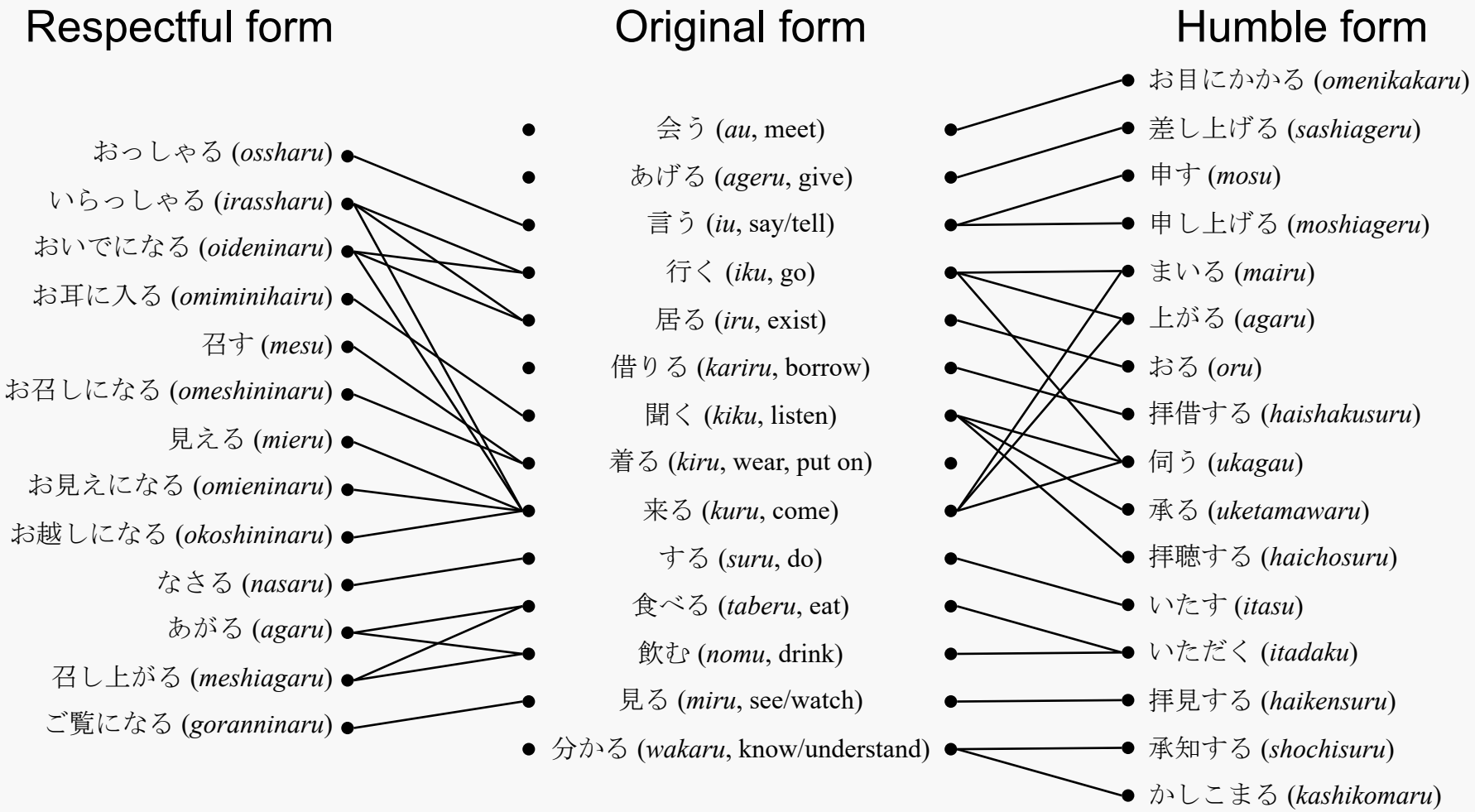
Wiktionary Edition

J-UniMorph

Verb from nouns are replaced by the single verb “*suru*”

Lexical Honorifics we use

- Morphemes do not correspond one-to-one with meanings



Randomly retrieve 10 entries (respectively)

Wiktionary Edition

Lemma	Inflected form	Hits
早寝	早寝できる	1,740
鼓する	鼓した	1,610
乗馬	乗馬し	87,600
脱会	脱会し	53,700
鯨飲	鯨飲します	708
断る	断らず	110,000
退部	退部さす	10
管理	管理します	51,600,000
放熱	放熱した	6,890
食指が動く	食指が動こう	369

J-UniMorph (Ours)

Lemma	Inflected form	Hits
話す	お話しください	824,000
渡す	渡させられた	1,220
休む	休まなかった	88,800
取る	取れます	14,500,000
貸す	貸させていただきます	2,500
走る	走らさない	19,100
乗る	乗らす	19,200
入れる	入れたがりません	219,000
聞く	承りません	9,230,000
撮る	撮らせません	4,720

Hits were assigned by us↑