NAACL 2024

The 21st SIGMORPHON workshop on Computational Morphology, Phonology, and Phonetics

Volume 1 - Proceedings of the Workshop

June 20, 2024

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-103-2

Introduction

Welcome to the 21st SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, to be held on June 20, 2024 as part of NAACL in Mexico City, Mexico. The workshop aims to bring together researchers interested in applying computational techniques to problems in morphology, phonology, and phonetics. Our program this year highlights the ongoing investigations into how neural and other learning models process phonology and word structure. We also publish work on a number of new datasets and resources in morphology.

We received 15 submissions, and after a competitive reviewing process, we accepted 9. The workshop is privileged to present two invited talks this year. Jian Zhu (University of British Columbia) and Naomi Feldman (University of Maryland) presented talks at this year's workshop.

Garrett Nicolai, Eleanor Chodroff, Çagri Çöltekin and Fred Mailhot, workshop organization team.

Organizing Committee

Co-Chair

Garrett Nicolai, University of British Columbia
Eleanor Chodroff, University of York
Ồagri Ồụltekin, University of Tựbingen
Fred Mailhot, Dialpad, Inc.

SIGMORPHON Officers

President: Garrett Nicolai, University of British Columbia
Secretary: Miikka Silfverberg, University of British Columbia
At Large: Eleanor Chodroff, University of York
At Large: Ồar Ồụltekin, University of Tựbingen
At Large: Fred Mailhot, Dialpad, Inc.

Program Committee

Reviewers

Brian Roark, Google Inc.
Aniello De Santo, University of Utah
Ekaterina Vylomova, University of Melbourne
Michael Ginn, University of Colorado
Kenneth Steimel, Cisco Systems Incorporated
Jelena Prokic, Leiden University
Nizar Habash, New York University Abu Dhabi
Khuyagbaatar Batsuren, National University of Mongolia
Sandra Kựbler, Indiana University
Changbing Yang, University of British Columbia
Adam Wiemerslage, University of Colorado Boulder
Kemal Oflazer, Carnegie Mellon University
Sarah Moeller, University of Florida
Cassandra L. Jacobs, University at Buffalo
Rob Malouf, San Diego State University
Nabil Hathout, CLLE, CNRS and Universitõ de Toulouse
Kate McCurdy, University of Edinburgh
Mathilde Hutin, Universitõ Paris-Saclay, CNRS, LIMSI
Micha Elsner, The Ohio State University
Daniel Dakota, Indiana University
Morgan Sonderegger, McGill University
Kristine Yu, University of Massachusetts Amherst
Ồar Ồụltekin, University of Tựbingen
Giorgio Magri, Centre National de la Recherche Scientifique
Indranil Dutta, Jadavpur University
Ewan Dunbar, University of Toronto

Keynote Talk
Invited Talk 1

Jian Zhu
University of British Columbia
2024-06-20 09:00:00 –

Abstract: Towards crosslinguistically generalizable speech technologies The diversity of human speech presents a formidable challenge to multilingual speech processing systems. Recently, accumulating evidence indicated that scaling up multilingual data and model parameters can tremendously improve the performance of multilingual speech processing. However, gathering large-scale data from every language in the world is an impossible mission. To tackle this challenge, my research group aims to develop multilingual speech processing systems that generalize to unseen and low-resource languages. Since most, if not all, human speech can be represented by around 150 phonetic symbols and diacritics, I argue that using International Phonetic Alphabet (IPA) as modeling units, rather than orthographic transcriptions, enables speech models to process and recognize sounds in unseen languages. In the past years, leveraging IPA, large-scale multilingual corpora and deep learning, my research team has built a series of massively multilingual speech datasets and technologies including multilingual grapheme-to-phoneme conversion, multilingual keyword spotting, multilingual forced alignment and multilingual phone recognition systems. In this talk, I will introduce our recent works towards crosslinguistically generalizable speech technologies and lessons we learned from working with a diversity of languages.

Bio: Jian Zhu is currently an assistant professor in the Linguistics Department at the University of British Columbia. He is primarily interested in developing multilingual speech and language technologies for low resource and zero resource languages. Trained as both a linguist and an engineer, he combines linguistic theories with data-driven methods in speech processing, natural language processing, network science and machine learning. Before that, he was a post-doctoral research fellow at Blablablab at the School of Information, University of Michigan. He obtained his Ph.D. in Linguistics and Scientific Computing from the Department of Linguistics and the Michigan Institute for Computational Discovery & Engineering at the University of Michigan.

Keynote Talk
Invited Talk 2

Naomi Feldman
University of Maryland
2024-06-20 14:30:00 –

Abstract: Modeling speech perception at scale Speech processing is a perfect test case for scaling up cognitive modeling. Recent advances in speech technology provide new tools that can be leveraged to better understand how human listeners perceive speech in naturalistic settings. At the same time, building cognitive models of human speech perception can highlight capabilities that are not yet captured by standard representation learning models in speech technology.

I begin by showing how incorporating unsupervised representation learning into cognitive models of speech perception can impact theories of early language acquisition. Infants' patterns of speech perception have traditionally been interpreted as evidence that they possess certain types of knowledge, such as phonetic categories (like 'r' and 'l') and representations of speech rhythm, but our cognitive modeling results point toward a different interpretation. If correct, this could radically change our view of how phonetic knowledge supports infants' acquisition of words and grammar, and could have broad implications for understanding the challenges associated with learning a new language in adulthood. I then outline ongoing work exploring the mechanisms that could support and, eventually, reproduce human listeners' ability to flexibly adapt to different accents and listening conditions. Together, these studies illustrate how speech representations can be optimized over short and long time scales to support robust speech processing.

This is joint work with Thomas Schatz, Yevgen Matusevych, Ruolan (Leslie) Famularo, Nika Jurov, Ali Aboelata, Grayson Wolf, Xuan-Nga Cao, Herman Kamper, William Idsardi, Emmanuel Dupoux, and Sharon Goldwater.

Bio: Naomi Feldman is an associate professor in the Department of Linguistics and the Institute for Advanced Computer Studies at the University of Maryland, where she is a member and former director of the Computational Linguistics and Information Processing (CLIP) Lab. Her research uses methods from machine learning and automatic speech recognition to formalize questions about how people learn and represent the structure of their language. She primarily uses these methods to study speech representations, modeling the cognitive processes that support learning and perception of speech sounds in the face of highly complex and variable linguistic input. She also computationally characterizes the strategies that facilitate language acquisition more generally, both from the perspective of learners, and from the perspective of clinicians.

Table of Contents

Program

Thursday, June 20, 2024

09:25 - 09:30 *Opening Remarks*

09:30 - 10:30 *Invited Talk 1":Jian Zhu*

10:30 - 11:00 *Break*

11:00 - 12:00 *Session 1*

J-UniMorph":Japanese Morphological Annotation through the Universal Feature Schema
Kosuke Matsuzaki, Masaya Taniguchi, Kentaro Inui and Keisuke Sakaguchi

Ye Olde French":Effect of Old and Middle French on SIGMORPHON-UniMorph Shared Task Data
William Kezerian, Lam An Wyner, Sandro Ansari and Kristine Yu

VeLePa":a Verbal Lexicon of Pame
Borja Herce

12:00 - 13:00 *Lunch*

13:00 - 14:00 *Session 2*

Acoustic barycenters as exemplar production targets
Frederic Mailhot and Cassandra L. Jacobs

Japanese Rule-based Grapheme-to-phoneme Conversion System and Multilingual Named Entity Dataset with International Phonetic Alphabet
Yuhi Matogawa, Yusuke Sakai, Taro Watanabe and Chihiro Taguchi

Decomposing Fusional Morphemes with Vector Embeddings
Michael Ginn and Alexis Palmer

14:00 - 15:00 *Invited Talk 2":Naomi Feldman*

Thursday, June 20, 2024 (continued)

15:00 - 15:30 *Session 3 (ACL Findings)*

Tokenization Matters":Navigating Data-Scarce Tokenization for Gender Inclusive Language Technologies
Anaelia Ovalle, Ninareh Mehrabi, Palash Goyal, Jwala Dhamala, Kai-Wei Chang, Richard Zemel, Aram Galstyan, Yuval Pinter and Rahul Gupta

Low-resource neural machine translation with morphological modeling
Antoine Nzeyimana

15:30 - 16:00 *Break*

16:00 - 17:00 *Session 4*

Different Tokenization Schemes Lead to Comparable Performance in Spanish Number Agreement
Catherine Arnett, Tyler Chang and Sean Trott

The Effect of Model Capacity and Script Diversity on Subword Tokenization for Sorani Kurdish
Ali Salehi and Cassandra L. Jacobs

More than Just Statistical Recurrence":Human and Machine Unsupervised Learning of Mori Word Segmentation across Morphological Processes
Ashvini Varatharaj and Simon Todd

VeLePa: a Verbal Lexicon of Pame

Borja Herce
University of Zurich

Abstract

This paper presents VeLePa, an inflected verbal lexicon of Central Pame (pbs, cent2154), an Otomanguean language from Mexico. This resource contains 12528 words in phonological form representing the complete inflectional paradigms of 216 verbs, supplemented with use frequencies. Inflected lexicons of non-WEIRD underresourced languages are urgently needed to expand digital capacities in these languages (e.g. in NLP). VeLePa contributes to this, and does so with data from a language which is morphologically extraordinary, with unusually high levels of irregularity and multiple conjugations at various loci within the word: prefixes, stems, tone, and suffixes constitute different albeit interrelated subsystems.

1 Introduction

Central Pame is an indigenous Mesoamerican language spoken by around 5000 people in and around Santa María Acapulco (San Luis Potosí, Mexico). The language is still acquired as a first language by children in various communities, but is endangered by the expansion of Spanish. The language lacks a standard written form, and extant documentation (e.g. Gibson & Bartholomew, 1979; Hurch; 2022) is insufficient, undigitized, and computationally largely unusable.

The language, however, like others in its family (e.g. Chichimec, see Palancar & Avelino, 2019; Herce, 2022) is a treasure trove of morphological complexity, due to the combination of the following two traits:

- Very high levels of irregularity, with many small inflection classes, many uniquely-behaving verbs, and a lot of suppletion.
- A morphological realization of subject and tense information which is distributed along the word into multiple inflectional layers: prefixes, tone, stem, and suffixes.

These properties make the system highly interesting and challenging to theoretical morphology as well as to NLP. Adding this language to databases like Unimorph (see McCarthy et al., 2020) and to morphological reinflection tasks would make these more representative of overall human language diversity and its limits.

2 Building VeLePa

To build an inflected lexicon of Central Pame verbs the first thing we need is language documentation. Although some inflectional paradigms were collected by SIL missionaries around 70 years ago (Gibson, 1950), these are insufficient in number and are hardly usable computationally due to inconsistencies.

Over the last four years, I have been documenting the language together with native speakers, mostly through the elicitation of inflected forms. For their orthographic transcription I adopt a phonemic approach, whereby only contrastive sounds are represented with different characters. International Phonetic Alphabet conventions are followed, as in the aforementioned previous work on the language. I thus avoid the problems of a Spanish-based orthography that is occasionally used to write the language locally and which does not represent features like vowel nasality, tone, consonant length, and the contrast between an mid-open and mid-closed front vowels.

The database (VeLePa) that is presented in this paper contains therefore the complete paradigms of a large number of verbs in phonological form. Every single one of the 12528 inflected forms that VeLePa contains (all 58 forms from 216 verbs) has been independently elicited (i.e. never extrapolated from other forms, as is often the case of these recources) and checked multiple times to avoid mistakes and inconsistencies (e.g. in the treatment of synonymous inflected forms like *dived*～*dove*). This is needed, first, because the language demands it. Most of the words in VeLePa (74%) have different forms, and syncretism (i.e. morphological whole-word identity) is never the result of different values being systematically the same across all lemmas as in other languages (e.g. English *do* INF, *do* 1SG.PRS, *do* 2SG.PRS, *do* 1PL.PRS, *do* 2PL.PRS, *do* 3PL.PRS).

Proceedings of the The 21st SIGMORPHON workshop on Computational Morphology, Phonology, and Phonetics, pages 1–6

Secondly, given the large degree of irregularity in the language, the linguist can almost never be sure to predict correctly one form of a verb from another. Eliciting every single form prevents underestimating complexity. At the same time, however, because VeLePa has been built with computational analysis in mind, cross-speaker and intra-speaker variability and free variation had to be ironed out in a way that this does not lead to an *over*estimation of morphological complexity. Although crucial, these types of quality controls are not always discussed and implemented in the compilation of inflected lexicons, particularly those from indigenous languages, as these tend to be produced by documentary linguists for whom the computational use of these resources is not a priority.

Given the absence of a standard of the language, and the unsuitability[1] of the orthography generally in use in the community, forms are represented in VeLePa in phonological transcription. Tones (High [H], Low [L], and Falling [F]) are indicated immediately after the (lowest) vowel of the syllable where they occur. Consonant gemination is indicated through a doubling of the corresponding consonant. To facilitate analysis, segmentations of prefix and stem have been included (indicated by "-"), as well as zero prefixes (indicated by "0"). These can be deleted if morphological decomposition is not needed. Other transcription choices are IPA-compliant. Typical forms are hence to-hoHʔo, 0-mbãLnʔ, laH-ppo, la-hõFlʔ, etc. or from a single verb la-nõH , ta-nõHn, ki-ŋõHik, 0-nõH, etc.

Every inflected form is tagged for its lemma (e.g. 'play') and morphosyntactic values (e.g. 1SG.PRS). As a further feature of interest to computational morphologists, for example those interested in the Paradigm Cell Filling Problem in a naturalistic setting, (see Ackerman et al., 2009; Blevins et al., 2017), I also provide a use frequency estimate of the different lemmas (see Figure 1, frequency estimated in number of tokens per million words) and morphosyntactic values (see Figure 2, frequency estimated as proportion of verbal tokens). These were derived from the frequency of forms in extant Central Pame texts (see Gibson et al, 1963; Gibson, 1966; Hurch 2022), and supplemented with subjective frequency estimates from native speakers (see Carrol, 1971) due to the small size of the available corpus (only 1171 verbal tokens) and its unbalanced thematic and genre composition.

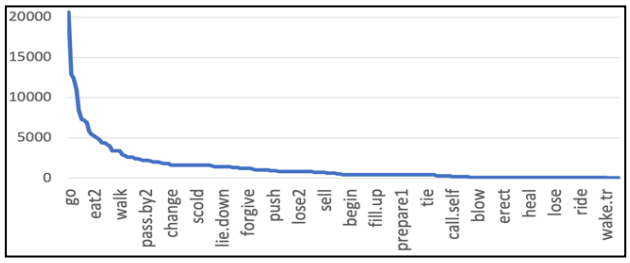

Figure 1: Frequency rankings of lemmas

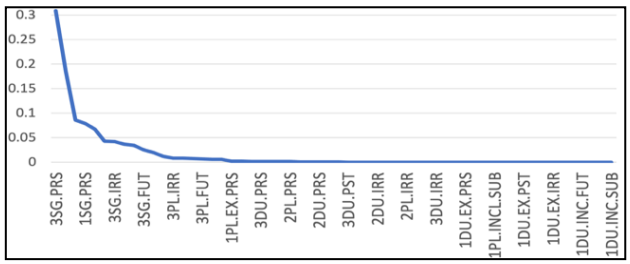

Figure 2: Frequency rankings of values

3 Analysis of system complexity

On the basis of VeLePa, freely available online at https://osf.io/xhyzm/?view_only=763f1c043e3f4c3787d0c93226e8b817, I analyze the morphological complexity and the predictability of the inflectional system as per the Paradigm Cell Filling Problem (see Ackerman et al., 2009). As mentioned in the introduction, one of the key idiosyncratic features of the language is the relative independence of prefixal, suffixal, tonal, and stem morphology. These four layers are analyzed separately below, through the following software:

- Qumín (Beniamine, 2018), for the automatic extraction of morphological alternations, and for the calculation of Information-Theoretic measures (e.g. conditional entropy of one form given another).
- Principal Parts Analyzer (Stump & Finkel, 2013), for the calculation of Set-Theoretic measures like the number of principal parts (i.e. the lowest number of forms required to predict the complete paradigm).

3.1 Prefixes

Despite their exhuberant allomorphy and the presence of stem-initial alternations, prefixes are straightforward to segment from stems. As the exemplary forms in Section 2 suggest, the prefix is the most changeable part of the word, and setting aside cases of zero-prefixed forms, corresponds generally to the first syllable of the word. Given this identification of prefixes, Pame verbs classify into 22 different

[1] While these are phonemic in the language, neither tone nor vowel nasalization nor consonant gemination are consistently represented in the traditional orthography.

inflectional classes, with a few comparatively frequent ones (see Table 1), and a long tail of (12) verbs which are prefixally unlike any other in the database.

type freq.	85	51	24	10	9	6	5	5
1SG.PRS	la	to	ti	la	la	ti	la	to
1DU.EX.PRS	ta	to	ti	ta	ta	ti	ta	to
1DU.INC.PR	ta	to	ti	ta	ta	ti	ta	to
1PL.EX.PRS	ta	to	ti	∅	ta	ti	wa	to
1PL.INC.PRS	ta	to	ti	∅	ta	ti	wa	to
2SG.PRS	ki	to	ti	ki	ki	ti	ki	la
2DU.PRS	ki	to	ti	ta	ki	ti	ta	la
2PL.PRS	ki	to	ti	∅	ki	ti	wa	la
3SG.PRS	wa	lo	li	∅	∅	li	∅	wa
3DU.PRS	wa	lo	li	∅	∅	li	∅	wa
3PL.PRS	∅	wa	ti	∅	∅	li	wa	wa

Table 1: Present prefixes of the 8 largest classes

As Table 1 shows,11 values of person-number are distinguished in the language, over 6 values of tense-aspect-mood. Due to the incompatibility of 1st and 3rd persons with the imperative mood, 58 values/cells exist in the Pame verb's paradigm. These fall into 39 areas of mutual interpredictability (see Table 2). These are those areas where the content of one cell (e.g. the 1PL.EX.PRS) allows to to predict that of another (e.g. 1PL.INC.PRS) and *vice versa*. In Pame this tends to mean their forms are always the same (e.g. ta/ta, to/to, ti/ti., ∅/∅, or wa/wa in Table 1).

	PRS	PST	IRR	SUB	FUT	IMP
1SG	1	9	16	24	32	-
1DU.EX	2		17	25		-
1DU.INC						-
1PL.EX	3	10	18	26	33	-
1PL.INC						-
2SG	4	11	19	27	34	37
2DU	5	12	20	28	32	38
2PL	6	13	21	29	33	39
3SG	7	14	22	30	35	-
3DU						-
3PL	8	15	23	31	36	-

Table 2: Prefix interpredictability areas

The average conditional entropy (i.e. a measure of the uncertainty involved in predicting one form from another) is 0.52 bits. On a different metric of complexity, 5 static principal parts are needed to predict the entire paradigm. These speak of the complexity of prefixal inflection in Central Pame, which is, however, lower than that of the other inflectional layers/subsystems in the language that will be presented in the next sections.

3.2 Stems

While all Pame verbs show prefixal and suffixal inflection, not all (96.3%) display stem alternation. Barring cases of suppletion, which occurs in twelve verbs, generally with different roots in SG/DU and PL, most of the morphological action in stems occurs on their consonantal onset. Sometimes, particularly in the 3PL across tenses, it involves the addition of segments, some other times it involves gemination, sometimes segmental changes, etc. These occur with somewhat recurrent distributions in the paradigm (see a summary of the largest classes in Table 3).

type freq.	16	13	6	5	5	5	5	5
1SG.PRS	pp	ʔ	ʔu	h	kk	pp	tt	tt
1DU.EX.PR	pp	ʔ	ʔu	h	kk	pp	tt	tt
1DU.INC.P	pp	ʔ	ʔu	h	kk	pp	tt	tt
1PL.EX.PRS	pp	ʔ	ʔu	h	kk	pp	tt	tt
1PL.INC.PR	pp	ʔ	ʔu	h	kk	pp	tt	tt
2SG.PRS	ppy	ʔy	ʔu	h	kky	pp	kky	kky
2DU.PRS	ppy	ʔy	ʔu	h	kky	pp	kky	kky
2PL.PRS	ppy	ʔy	ʔu	h	kky	pp	kky	kky
3SG.PRS	pp	ʔ	ʔu	h	kk	pp	tt	tt
3DU.PRS	pp	ʔ	ʔu	h	kk	pp	tt	tt
3PL.PRS	b	lʔ	tʔ	th	kh	pp	lh	lʔ
1SG.PST	w	ʔu	ʔu	h	ku	pp	t	t

Table 3: Present stem onsets of the 8 largest classes

Given the regularities in the distribution over values of different alternations, the 58 cells of the Pame verb paradigm are grouped into 29 interpredictability areas (see Table 4). The average conditional entropies between them is 0.63 bits, and 6 principal parts are minimally needed to be able to predict the complete stem paradigm without uncertainty.

	PRS	PST	IRR	SUB	FUT	IMP
1SG	1	7	13		22	-
1DU.EX						
1DU.INC						
1PL.EX	2	8	14	19	23	
1PL.INC						
2SG	3	9	15		24	28
2DU						
2PL	4	10	16	20	25	29
3SG	5	11	17	21	26	-
3DU						
3PL	6	12	18		27	

Table 4: Stem interpredictability areas

3.3 Tones

Tone (high, falling, or low) occurs in Pame in the stressed syllable, which can be either the final one (i.e. the root), or the penultimate (i.e. the prefix). Tone and stress are further intertwined in the language in that only the high tone occurs when the stressed syllable is the penultimate. The result is that only 4 tone-stress profiles are possible in any given word.

While all or most Pame verbs are inflectable in the other morphological layers, tone is different in that most verbs (66.2%) have a single tone across the paradigm (see the 4 largest classes in Table 5). Despite this, the PCFP is a considerable challenge because there is no way to predict, from the tonal value of a given form, whether this same tone will be found across the paradigm or in specific domains only, of which 19 exist (see Table 6).

type freq.	52	47	25	18	8	5	5	4
3SG.PRS	-L	-H	H-	-F	-L	H-	-H	-H
3DU.PRS	-L	-H	H-	-F	-L	H-	-H	-H
3PL.PRS	-L	-H	H-	-F	-L	-H	-H	-H
1SG.PST	-L	-H	H-	-F	-L	-H	-L	-H
1DU.EX.PST	-L	-H	H-	-F	-L	-H	-L	-H
1DU.INC.PS	-L	-H	H-	-F	-L	-H	-L	-H
1PL.EX.PST	-L	-H	H-	-F	-L	-H	-L	-H
1PL.INC.PST	-L	-H	H-	-F	-L	-H	-L	-H
2SG.PST	-L	-H	H-	-F	-F	-H	-F	-L
2DU.PST	-L	-H	H-	-F	-F	-H	-F	-L
2PL.PST	-L	-H	H-	-F	-F	-H	-F	-L

Table 5: Tones of the 8 largest classes

<table>
<tr><th></th><th>PRS</th><th>PST</th><th>IRR</th><th>SUB</th><th>FUT</th><th>IMP</th></tr>
<tr><td>1SG</td><td rowspan="3">1</td><td rowspan="3">6</td><td rowspan="3">10</td><td rowspan="3"></td><td rowspan="3">16</td><td rowspan="5">-</td></tr>
<tr><td>1DU.EX</td></tr>
<tr><td>1DU.INC</td></tr>
<tr><td>1PL.EX</td><td rowspan="2">2</td><td rowspan="2">7</td><td rowspan="2">11</td><td rowspan="2"></td><td rowspan="2">17</td></tr>
<tr><td>1PL.INC</td></tr>
<tr><td>2SG</td><td rowspan="2">3</td><td rowspan="2">8</td><td rowspan="2">12</td><td rowspan="2"></td><td rowspan="2">18</td><td rowspan="2">12</td></tr>
<tr><td>2DU</td></tr>
<tr><td>2PL</td><td>4</td><td>9</td><td>13</td><td></td><td>19</td><td>13</td></tr>
<tr><td>3SG</td><td rowspan="2">1</td><td rowspan="2">6</td><td rowspan="2">14</td><td rowspan="2"></td><td rowspan="2"></td><td rowspan="3">-</td></tr>
<tr><td>3DU</td></tr>
<tr><td>3PL</td><td>5</td><td>7</td><td>15</td><td></td><td></td></tr>
</table>

Table 6: Tone interpredictability areas

Despite the small number of possible values of tone, the average conditional entropy between these domains is 1.01, and one would need minimally 7 principal parts to be able to predict with certainty the tone of every inflected form. These values are the highest among all four inflectional layers.

3.4 Suffixes

While prefixes, stems, and tones encode, often redundantly, different values of subject person-number, and tense-aspect-mood, suffixes tend to encode person-number almost exclusively. Pame suffixes are always non-syllabic, attaching as a syllable coda when the stem finishes in a vowel (e.g. kowwaL +i > kowwaLi; kowwaL +nʔ > kowwaLnʔ) but modifying the stem ending when the root already has a coda (e.g. toŋgoãHn +i > toŋgoãHiŋ, toŋgoãHn +nʔ > toŋgoãHnʔ). This gives rise to unpredictability in that, given a suffixed form (e.g. one which contains an underlying suffix -nʔ), it cannot be known what the unsuffixed form is (e.g. Ø vs -n in the verbs above).

Alongside this source of maybe "superficial" unpredictability, suffixes also change from verb to verb. As the forms in Table 7 show, some have a 2DU suffix -k while others do not, and some have a 3PL suffix -t while others do not. Mainly these two sources of unpredictability combine to generate a PCFP challenge comparable to the other inflectional layers, with 14 areas of interpredictability (see Table 8), 0.62 bits of average conditional entropy, and 6 static principal parts.

type freq.	50	27	22	8	8	6	6	6
1SG.PRS	Ø	ŋ	ʔ	n	Ø	ʔ	ŋ	t
1DU.EX.PR	mʔ	mʔ	mʔ	nʔ	mʔ	mʔ	mʔ	nʔ
1DU.INC.PR	Ø	Ø	ʔ	ŋ	Ø	ʔ	Ø	t
1PL.EX.PRS	nʔ	nʔ	nʔ	nʔ	nʔ	nʔ	nʔ	nʔ
1PL.INC.PR	n	n	n	n	n	n	n	n
2SG.PRS	Ø	ŋ	ʔ	n	Ø	ʔ	ŋ	t
2DU.PRS	Ø	Ø	ʔ	ŋ	k	ʔk	Ø	t
2PL.PRS	n	n	nʔ	n	n	nʔ	n	n
3SG.PRS	Ø	ŋ	ʔ	n	Ø	ʔ	ŋ	t
3DU.PRS	Ø	Ø	ʔ	ŋ	Ø	ʔ	Ø	t
3PL.PRS	Ø	ŋ	ʔ	n	t	ʔ	nt	t

Table 7: Suffixes of the 8 largest classes

<table>
<tr><th></th><th>PRS</th><th>PST</th><th>IRR</th><th>SUB</th><th>FUT</th><th>IMP</th></tr>
<tr><td>1SG</td><td>1</td><td>10</td><td colspan="2">1</td><td></td><td>-</td></tr>
<tr><td>1DU.E</td><td colspan="5">2</td><td>-</td></tr>
<tr><td>1DU.IN</td><td>3</td><td>11</td><td colspan="2">3</td><td></td><td>-</td></tr>
<tr><td>1PL.EX</td><td colspan="5">4</td><td>-</td></tr>
<tr><td>1PL.IN</td><td colspan="5">5</td><td>-</td></tr>
<tr><td>2SG</td><td>6</td><td>12</td><td colspan="2">6</td><td></td><td></td></tr>
<tr><td>2DU</td><td>7</td><td></td><td colspan="4">7</td></tr>
<tr><td>2PL</td><td colspan="6">8</td></tr>
<tr><td>3SG</td><td>1</td><td>10</td><td colspan="2">1</td><td></td><td>-</td></tr>
<tr><td>3DU</td><td>3</td><td>11</td><td colspan="2">3</td><td></td><td>-</td></tr>
<tr><td>3PL</td><td colspan="5">9</td><td>-</td></tr>
</table>

Table 8: Suffixal interpredictability areas

4 Discussion

An inflectional system with the complexity of any of these layers would be considered quite complex. The (in)famous Latin verbs, for example, have 4 principal parts, 0.28 bits of average conditional entropies, and 15 zones of interpredictability (see Pellegrini, 2020), yet this is almost consistently simpler than any of the inflectional subsystems that coexist within Pame verbs. The overall system, hence, would appear to test the very limits of human linguistic cognition. How do speakers manage to successfully learn and use a system like this? The answer might lie in predictability *between* inflectional layers. While that between cells is explored more often, as I have done in the previous Section 3, this does not mean that predictability between different slots or properties of a single word plays no role. Preliminary assessment of how much information one layer provides about another in Pame can be obtained from Normalized Mutual Information (NMU), calculated through the R package aricode (Chiquet et al., 2020). Results in Table 9 show NMU oscillates between 0.18 and 0.56, which means the lexical classifications of different layers are highly informative about each other. To mention a few examples, the largest prefixal class is close to incompatible with the absence of stem alternation, the second and third largest prefixal classes are incompatible with tonal alternations, etc.

	tone-stress	stems	suffixes
prefixes	0.261	0.558	0.346
tone-stress		0.270	0.179
stems			0.250

Table 9: NMU between the different slots

Beyond these between-layer predictive relations, another challenging aspect of Pame verb morphology is the unsystematic nature of syncretism. While this is not infrequent in the language (26% of forms), this does not occur systematically, in that there are no cells in the paradigm that are always syncretic. It is remarkable, for example, that prefixal inflection classes (see the largest ones in Table 1) differ not only in their use of different allomorphs, but also in their partition of the semantic space. Because the pattern of contrasts is different in every class of verbs, it must make the Paradigm Cell Finding Problem (see Boyé & Schalchli, 2019) extremely challenging.

A final challenge that Pame verbs present is what Erdmann et al. (2020) have called the Paradigm Identification Problem. Given the amount of suppletion, stem alternation and allomorphy in the system, predicting the lemma and morphosyntactic value of a form from its morphology must also be complicated. The same markers are reused with different functions in different verbs classes (As show in Table 1, for example, la- occurs in the 1SG.PRS prefix in some verbs but as the 2.PRS in other verbs, wa- occurs as a 3SG/DU.PRS in some verbs, but as 3PL.PRS in others, or as 3.PRS or 3.PL in others, etc. These and other aspects can now be explored computationally through the resource VeLePa.

5 Conclusion

This paper has reported on the compilation of a Verbal Lexicon of Pame (VeLePa), specifically with computational applications in mind. It has also presented some preliminary quantitative analyses of this inflectional system around the topics of the Paradigm Cell Filling Problem, and related challenges that speakers and learners of a language face when using and/or acquiring inflectional morphological patterns.

This system, like other Otomanguean ones (see e.g. Cruz et al. 2020) is remarkable because its morphology deviates very significantly and in several dimensions from the canonical (Corbett 2009) most straightforward one. It is structured into several morphological slots which work together (see the phenomena of Multiple, Distributed or Extended Exponence, Harris 2017) into expressing tense-aspect-mood and subject person-number values. Each slot is, furthermore, organized into a large number of inflection classes and contains multiple isolated irregularities.

VeLePa is expected to contribute to both theoretical linguistic analysis and to NLP, allowing the inclusion (e.g. into reinflection tasks) of a language that is both highly complex and typologically very different from the better documented (Indo-)European ones.

Acknowledgments

Thanks are due to the Swiss Society for Endangered Languages and to the University of Zurich's GRC for financial support to conduct the fieldwork on which this research is based. I also thank the patient help of my main informants Aniceto and Josefa. All mistakes are, of course, mine.

References

Ackerman, F., Blevins, J. P. and Malouf, R. (2009). Parts and wholes: Patterns of relatedness in complex morphological systems and why they matter. In J. P. Blevins and J. Blevins (Eds.), Analogy in grammar: Form and acquisition: 54–82. Oxford: OUP.

Beniamine, S. (2018). Classifications flexionnelles. Étude quantitative des structures de paradigmes. PhD diss., Université Sorbonne Paris Cité.

Blevins, J. P., Milin, P. and Ramscar, M. (2017). The Zipfian paradigm cell filling problem. In Perspectives on morphological organization: 139-158. Brill.

Boyé, G., and Schalchli, G. (2019). Realistic data and paradigms: The paradigm cell finding problem. Morphology, 29: 199-248.

Carroll, J. B. (1971). Measurement properties of subjective magnitude estimates of word frequency. Verbal Learning & Verbal Behavior, 10: 722-729.

Chiquet, J., Rigaill, G., Sundqvist, M., Dervieux, V., and Bersani, F. (2020). Package 'aricode'.

Corbett, G. G. (2009). Canonical inflectional classes. In *Selected proceedings of the 6th Décembrettes: Morphology in Bordeaux*:1-11.

Cruz, H., Stump, G., and Anastasopoulos, A. (2020). A resource for studying chatino verbal morphology. *arXiv preprint arXiv:2004.02083*.

Erdmann, A., Elsner, M., Wu, S., Cotterell, R., and Habash N. (2020). The Paradigm Discovery Problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*: 7778–7790.

Gibson, L. F. (1950). Verb paradigms in Pame. https://www.sil.org/resources/archives/53023.

Gibson, L. F. (1966). The Man Who Abandoned his Wife. A Pame Story. Tlalocan 5, no. 2: 169-177.

Gibson, L. F., Olson, D. and Olson, A. (1963). Four Pame Texts. Tlalocan 4, no. 2: 125-143.

Harris, A. C. (2017). *Multiple exponence*. Oxford University Press.

Herce, B. (2022). Possessive inflection in Chichimec inalienable nouns: The morphological organization of a closed irregular class. Studies in Lang. 46, 4: 901-933.

Hurch, B. (2022). Pame (central) de Santa María Acapulco, Santa Catarina, San Luis Potosí. In Y. Lastra (Ed.) Archivo de Lenguas Indígenas de México. México DF: El Colegio de México.

McCarthy, A. D., Kirov, C., Grella, M., Nidhi, A., Xia, P., Gorman, K., Vylomova, E. et al. (2020). UniMorph 3.0: Universal Morphology. In Proceedings of The 12th LREC: 3922-3931. ELRA.

Palancar, E. L. and Avelino, H. (2019). Inflectional complexity and verb classes in Chichimec. Amerindia 41: 323-360.

Pellegrini, M. (2020). Using LatInfLexi for an entropy-based assessment of predictability in Latin inflection. In Proceedings of LT4HALA 2020: 37-46.

Stump, G. and Finkel, R. A. (2013). Morphological typology: From word to paradigm. Cambridge: CUP

J-UniMorph: 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-UniMorph, developed based on the UniMorph feature schema. This dataset addresses the unique and rich verb forms characteristic of the language's agglutinative nature. J-UniMorph 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-UniMorph 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-UniMorph, comparing it with the Wiktionary Edition. We will release J-UniMorph 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-UniMorph. 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-UniMorph 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-UniMorph (§2). We then explain the data creation process (§3). As illustrated in Figure 1, this

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

Proceedings of the The 21st SIGMORPHON workshop on Computational Morphology, Phonology, and Phonetics, pages 7–19

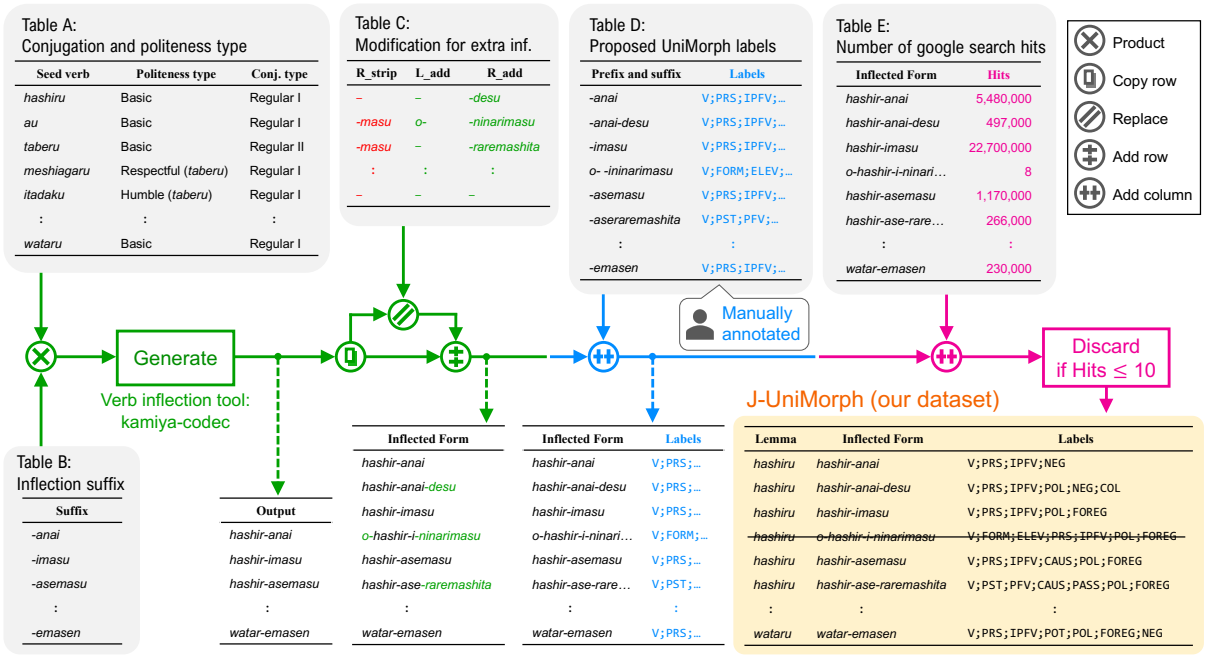

Figure 1: Overview of the J-UNIMORPH creation process: First, we generate inflected forms from seed verbs (Table A, detailed in §3.1) and inflection suffix (Table B, detailed in §3.2) using the verb inflection tool, `kamiya-codec`. This is followed by modifying and adding inflected forms that the tool 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 UniMorph 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 provides examples of Regular I and II verbs.

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

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 UniMorph 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.

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

行く/ *iku*	行く/ *iku*
行かれる/ *ika-reru*	いらっしゃる/ *irassharu*
V;PRS;IPFV;**ELEV**	V;**FORM**;**ELEV**;PRS;IPFV
会う/ *au*	会う/ *au*
会われる/ *awa-reru*	お会いになる/ *o-ai-ninaru*
V;PRS;IPFV;**ELEV**	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

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/

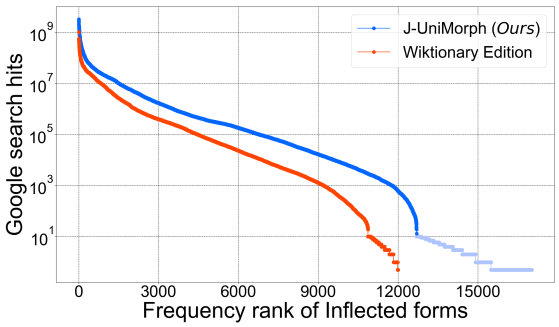

Figure 2: The relationship between the frequency rank of inflected forms and their corresponding number of Google search hits, highlighting a long-tail distribution pattern, regarding J-UNIMORPH and Wiktionary Edition, respectively. Both graphs exhibit a clear trend shift when the number of hits falls 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 UniMorph 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*" (≈ 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**.*
≈ 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**.*
≈ 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.

[16] https://github.com/cl-tohoku/J-UniMorph

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 Conference*, 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

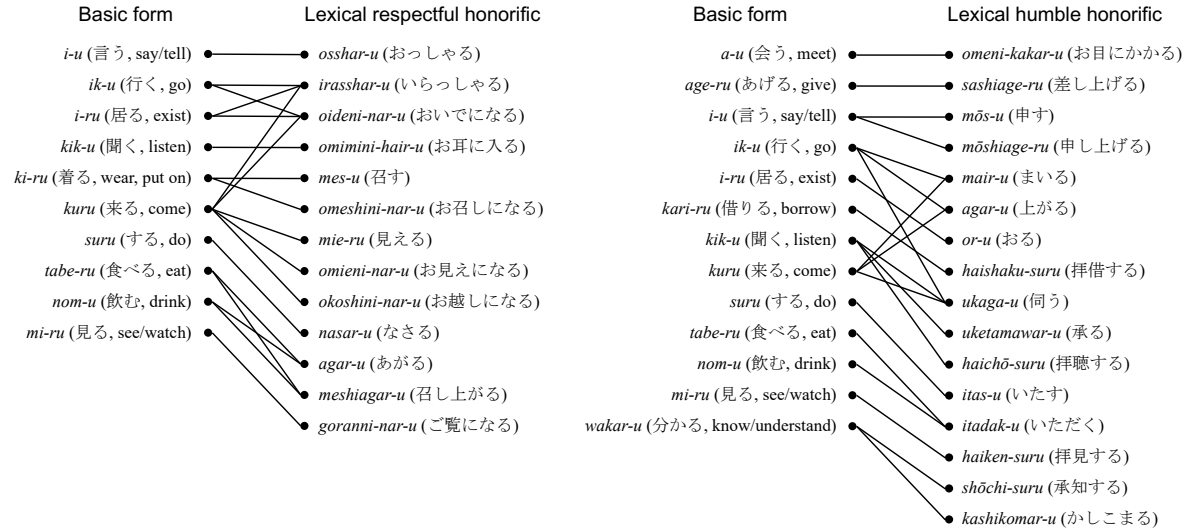

Figure 4: Correspondence between the basic forms and the lexical honorifics adopted in J-UNIMORPH.

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-UNIMORPH

We provide several details on the excluded forms in J-UNIMORPH, 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-UNIMORPH due to their morphological equivalence to the subsidiary verbs that are already incorporated into J-UNIMORPH 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-UNIMORPH 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-UNIMORPH.

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-UNIMORPH.

[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-**i**ru* /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.

More than Just Statistical Recurrence: Human and Machine Unsupervised Learning of Māori Word Segmentation across Morphological Processes

Ashvini Varatharaj
Department of Linguistics,
University of California Santa Barbara
ashvinivaratharaj@ucsb.edu

Simon Todd
Department of Linguistics,
University of California Santa Barbara
& NZILBB, University of Canterbury
sjtodd@ucsb.edu

Abstract

Non-Māori-speaking New Zealanders (NMS) are able to segment Māori words in a highly similar way to fluent speakers (Panther et al., 2024). This ability is assumed to derive through the identification and extraction of statistically recurrent forms. We examine this assumption by asking how NMS segmentations compare to those produced by Morfessor, an unsupervised machine learning model that operates based on statistical recurrence, across words formed by a variety of morphological processes. Both NMS and Morfessor succeed in segmenting words formed by concatenative processes (compounding and affixation without allomorphy), but NMS also succeed for words that invoke templates (reduplication and allomorphy) and other cues to morphological structure, implying that their learning process is sensitive to more than just statistical recurrence.

1 Introduction

Humans have a powerful ability to build implicit linguistic knowledge incidentally, based on passive processes that identify and extract statistically recurrent patterns (Saffran et al., 1996; Frank et al., 2013; Aslin, 2017). For example, New Zealanders who are regularly ambiently exposed to Māori, but do not speak it, nevertheless have Māori lexical and phonotactic knowledge (Oh et al., 2020; Panther et al., 2023) and can morphologically segment Māori words at above-chance levels (Panther et al., 2024). These findings imply that regular exposure to a language yields a *proto-lexicon*: an implicit memory-store of forms that recur with statistical regularity in the language, including both words and word-parts (Ngon et al., 2013; Johnson, 2016).

In this paper, we are concerned with the way that the proto-lexicon is constructed, and the way that its construction interacts with language structure. We examine the extent to which the ability of non-Māori-speaking New Zealanders (NMS) to morphologically segment Māori words is explained by naive statistical learning, in which their proto-lexicon is assumed to be formed purely through the identification and extraction of statistically recurrent forms in ambient Māori. To do so, we generate expectations for what morphological segmentation would look like through naive statistical learning processes from Morfessor (Creutz and Lagus, 2007; Virpioja et al., 2013), an unsupervised Bayesian segmentation model. We compare the segmentations produced by Morfessor to those produced by NMS and examine how they vary across words formed by different morphological processes.

Through two analyses, we argue that NMS do more than a naive statistical learning model would suggest. First, we compare the segmentations of Morfessor and NMS across Māori words formed by affixation and compounding, both concatenative processes, and words formed by reduplication, a templatic process. We find that both are accurate on words formed by affixation and compounding, but NMS are more accurate on words formed by reduplication, suggesting that NMS identify and extract both statistically recurrent forms and higher-level abstract templates. Then, zooming in on words formed by concatenative processes, we ask whether there are other cues to morphological structure that NMS may be picking up on, such as vowel length. We compare the performance of Morfessor across real Māori words that may contain such cues and constructed words that have the same statistical properties but lack any reliable alternative cues to morphological structure. We find that Morfessor is worse at segmenting real words, suggesting that successful learning by NMS requires sensitivity to more cues than just statistical recurrence.

2 Background

2.1 Statistical learning of language

How humans learn to extract knowledge from their environment is one of the fundamental questions in

Proceedings of the The 21st SIGMORPHON workshop on Computational Morphology, Phonology, and Phonetics, pages 20–31

cognitive science. Implicit learning – the process of learning without intention, and even without the awareness of what has been learned (Williams, 2020) – is one of the main ways we learn from our surroundings. Implicit learning underlies various essential skills such as language comprehension and production, intuitive decision making, and social interaction (Rebuschat, 2015). A particularly prominent form of implicit learning is *statistical learning*[1]. Statistical learning refers to the process of extracting statistical regularities from input and adapting to them, based on considerations of frequency, variability, distribution, and co-occurrence (Saffran et al., 1996). Humans are highly sensitive to such statistical regularities and implicitly learn them from birth (Bulf et al., 2011; Gervain et al., 2008; Teinonen et al., 2009).

While most work on statistical learning has focused on studying infants (Saffran, 2001; Pelucchi et al., 2009) in lab-based setups, recent works have shown that adults are also capable of statistical learning of implicit linguistic knowledge through everyday exposure to a language they don't speak. Non-speakers of Māori in New Zealand (Oh et al., 2020; Panther et al., 2023) and Spanish in California and Texas (Todd et al., 2023) show evidence of implicit phonotactic and lexical knowledge of their respective ambient languages. However, this knowledge appears to be weaker in the case of Spanish than in Māori, and it has been argued that this difference may partly derive from differences in morphological structure (Todd et al., 2023).

In addition to having implicit phonotactic and lexical knowledge of Māori, non-Māori-speaking New Zealanders (NMS) can morphologically segment Māori words in a highly similar way to fluent speakers (Panther et al., 2024). This ability is facilitated by their possession of a *proto-lexicon* (Johnson, 2016; Ngon et al., 2013), a large implicit memory-store of the forms of words and word-parts that recur with statistical regularity in the language, called *morphs*. These morphs are defined by form, without consideration of meaning; thus, they may or may not correspond to underlying morphemes, and may even include phonological sequences that span word boundaries as long as they are statistically recurrent in the language (Ngon et al., 2013).

[1]While early literature on statistical learning focused narrowly on phonotactic transition probabilities, in this work we use the term more broadly to refer to the learning of any statistical properties of language.

2.2 Morphological segmentation

Many modern approaches to morphological segmentation use supervised learning, independently or in combination with unsupervised learning (e.g., Rouhe et al., 2022). In this work, we are attempting to model human learning of morphological segmentation that occurs without explicit instruction. For this reason, we use unsupervised learning.

Unsupervised morphological segmentation provides us an avenue to simulate implicit statistical learning processes. In this work, we use Morfessor Baseline (Creutz and Lagus, 2007; Virpioja et al., 2013), a popular unsupervised morphological segmentation model with an underlying generative process that is very simple and highly compatible with a naive model of statistical learning of morphological structure. Morfessor identifies a set of statistically recurrent morphs under the assumption that words are formed through the concatenation of these morphs, without phonological alternations, and without constraints applied to positioning, sequencing or morphosyntactic category.

Morfessor identifies the set of statistically recurrent morphs, which it calls a *lexicon* (and which is analogous to a human proto-lexicon), using a Minimum Description Length framework (Rissanen, 1978). This lexicon is therefore the smallest set of simplest morphs that can be combined to generate the training data with highest probability. The lexicon is constructed dynamically through several passes over the training data, where the cost of adding a morph to the lexicon at any point is based on the morph's complexity and its frequency of recurrence across the words segmented so far.

While Morfessor's assumptions are simple, there are simpler models that have gained currency recently as tokenizers in Natural Language Processing (e.g., Sennrich et al., 2015; Kudo, 2018; Wu et al., 2016). Like Morfessor, these models identify a set of morphs (which they call *subwords*) that generate the training data with highest probability, assuming only simple concatenation. However, unlike Morfessor, they require the number of morphs to be predetermined, and they do not simultaneously consider the complexity of proposed morphs, which we consider to be important for our modeling of human learning. There are also many morphological segmentation models that are more complex than Morfessor, such as Adaptor Grammars (Johnson and Griffiths, 2007; Eskander et al., 2016; Godard et al., 2018). These models offer fine-

grained assumptions about precisely how morphs may be combined, in contrast to Morfessor's assumption of simple concatenation. It is the relative simplicity of Morfessor that makes it a suitable baseline model of idealized statistical learning of a proto-lexicon, especially in a language that uses primarily concatenative morphological processes.

Morfessor's statistical learning approach mirrors that which has been assumed for NMS (Oh et al., 2020). Both are learning to segment based on statistical patterns in the language they are exposed to, without getting feedback. In both cases, the learners are identifying recurring forms and extracting them as morphs in a (proto-)lexicon. By using Morfessor as a baseline of comparison for NMS, we can understand how much of NMS' implicit knowledge is due to simple statistical learning processes. We expect Morfessor to perform best with words formed by concatenative morphological processes and to struggle with words formed by other morphological processes that are beyond the scope of its simple assumptions; if NMS do not struggle in the same way, then we may infer that they are doing more than just tracking statistical recurrence as Morfessor would assume.

2.3 The Māori Language

The Māori language consists of ten consonants <p, t, k, m, n, ng, w, r, wh, h>, five short vowels <a, e, i, o, u>, and five long vowels <ā, ē, ī, ō, ū>. The orthographic system is highly transparent: each grapheme or digraph corresponds to a unique phoneme. The basic timing unit is the mora, where short vowels count as one mora each and long vowels count as two (Harlow, 2007). The syllable structure is (C)V(V), but is often treated as (C)V for modeling purposes because of the complexity of distinguishing diphthongs from sequences of monopthongs (Bauer, 1993; Oh et al., 2020). There is a general minimality constraint which states that (content) words and morphs consist of at least two moras (Bauer, 1993, p. 544), and it has been argued that words consisting of four or more moras are highly likely to be morphologically complex (Krupa, 1968; de Lacy, 2003).

There are three main morphological processes in Māori: reduplication, affixation, and compounding (Bauer, 1993; Harlow, 2007). Reduplication consists of the repetition of part of a base, following one of many templates (see e.g. Keegan, 1996; Todd et al., 2022). Because of this reliance on a template, we refer to reduplication as a *templatic* process.[2] Affixation and compounding both consist of the concatenation of morphs that need not have any relation to each other in form, and thus we refer to them as *concatenative* processes. At a distributional level, affixation and compounding are distinguished by the fact that affixation causes a small set of four (Bauer and Bauer, 2012) or five (Harlow, 2007) productive morphs[3] to recur across many words, whereas compounding causes a large set of morphs to each recur across relatively fewer words (Bauer, 1993, p. 519).

Māori morphophonology may be described as strictly local: there are no morphophonological alternations, no phonologically discontiguous morphemes, and no long-distance phonological dependencies. However, there is affix allomorphy, in which affixes follow phonological templates, with different thematic consonants that are to some extent predictable (Parker Jones, 2008). This allomorphy is restricted to the passive and nominalizing suffixes, each of which has default and non-default allomorphs that are or are not consistent with major phonological templates (passive: *-Cia*; nominal: *-Canga*; both for thematic consonant *C*).

At a high level, the strictly local nature of Māori morphophonology accords exactly with the assumptions of Morfessor. However, the templates that underpin reduplication and affix allomorphy are not accounted for by Morfessor's underlying generative model. This means that the three morphological processes in Māori are consistent with Morfessor's assumptions to different extents, which allows us to examine how the degree to which Morfessor reflects NMS morphological segmentations is affected by morphological structure.

3 Analysis 1: Sensitivity to templates

Our first analysis examines Morfessor and NMS segmentations of Māori words formed through different morphological processes. For each learner, we identify the sensitivity to general templates and the importance of morphological concatenativity by comparing segmentation performance across words formed by reduplication and words formed by affix-

[2] We avoid the label templatic *morphology* so as to avoid confusion with root-and-pattern morphology such as is found in Semitic languages.

[3] Whether there are held to be four or five productive affixes depends on where the analyst draws the line between affixation and phrasal constructions. It is not entirely straightforward to designate these affixes as clearly inflectional or clearly derivational (Bauer and Bauer, 2012).

ation or compounding (Section 3.2), as well across cases of affixation that follow salient allomorphic templates to different extents (Section 3.3). This analysis reveals how unsupervised learning of morphological segmentation is sensitive to linguistic structure, and the extent to which the underlying assumptions of Morfessor make it a plausible model of naive statistical learning of morphological segmentation in humans.

3.1 Data

The analysis is conducted over a subset of words from the stimuli of Panther et al. (2024), which we aggregated into categories based on the morphological processes they likely represent (described below). We used the segmentations provided by a fluent Māori speaker (MS), collected by Oh et al. (2020), as a gold standard. To ensure that the morphological processes assumed by our categorizations adequately reflect those revealed by the MS segmentations, we filtered each category to only include words in which the MS segmentation is consistent with the assumed morphological process. After this filtering, the analysis is based on 3,919 words, categorized as follows:

Monomorphemic: Words consisting of 2 or 3 moras ($N = 622$ / 295, respectively) that did not receive any boundaries in the MS segmentation.

Reduplication: Words that were segmented by the MS in a manner consistent with one of four reduplication templates[4]: total (e.g., *paki+paki*; $N = 439$), right (e.g., *tākai+kai*; $N = 276$), left (e.g., *nu+nui*; $N = 111$), or left with lengthening (e.g., *kā+kahu*; $N = 36$). Total reduplication is the most salient of these templates.

Affixation: Words in which the MS recognized either the causative prefix *whaka-* ($N = 296$), a passive suffix ($N = 437$), or a nominalizing suffix ($N = 203$). The suffixes have many allomorphs which differ in terms of frequency and consistency with a major phonological template (passive: *-Cia*; nominal: *-Canga*; both for thematic consonant *C*), including, in descending order of frequency: template-consistent defaults (passive: *-tia*, *-hia*, *ngia*; nominal: *-tanga*, *-hanga*)[5]; non-template-consistent defaults (passive: *-a*; nominal: *-nga*[6]); template-consistent non-defaults (passive: *-kia*, *-mia*, *-ria*, *-whia*; nominal: *-kanga*, *-manga*, *-ranga*, *-whanga*); and non-template-consistent non-defaults (passive: *-ia*, *-na*, *-nga*[6], *-ina*, *-hina*, *-kina*, *-whina*; nominal: *-anga*).

Compounding: Words that consist of four or more moras, without reduplication or affixation, and for which the MS identified at least one boundary ($N = 1204$; a subset of the 'polymoraics' explored by Panther et al., 2024).

For each word, we compare the gold standard segmentation provided by the MS to the segmentations provided by Morfessor and NMS. The Morfessor segmentations were obtained from a model trained with default settings (using the implementation of Virpioja et al., 2013) on 19,595 word types from the Te Aka dictionary (Moorfield, 2011). The NMS segmentations are based on data collected by Panther et al. (2024) in a word-splitting task, where NMS participants split orthographically-presented words into pieces by placing any number of boundaries at any site between two letters.[7] To aggregate segmentations of a single word across participants, we used a majority-vote approach: we coded each site as containing a boundary if and only if the majority of participants who responded to that word placed a boundary there.

3.2 Analysis 1A: Morphological processes

We first analyze the degree to which segmentations by Morfessor and NMS match the gold standard segmentations, across categories of words formed by different morphological processes. We examine variation across categories, as well as how this variation differs between learners.

3.2.1 Methods

There are many metrics that compare a learner's morphological segmentations to a gold standard (Virpioja et al., 2011). We use the simple metric of boundary precision and recall, which considers

[4]The reduplication category includes some cases where there is both reduplication and compounding. We assess the placement of all boundaries in such cases, regardless of whether they separate the reduplicant from the base or one compound component from another.

[5]Dialects differ in terms of whether the default thematic consonant is <t>, <h>, or <ng>, though it is most commonly <t> (Harlow, 2007).

[6]*-nga* is both a passive suffix and a nominalizing suffix. As a passive suffix, it is not a default allomorph, but as a nominalizing suffix, it is. Our analysis of *-nga* is restricted to its occurrence as a nominalizing suffix.

[7]We analyze the same filtered subset of NMS participants as Panther et al. (2024): 195 individuals who have lived in NZ since the age of 7, have never taken any linguistics courses, and have explicit knowledge of few Māori words and grammatical structures. For full details of the experiment design and filtering criteria, see Panther et al. (2024).

Table 1: Macro-averaged precision and recall for Morfessor and NMS across categories of words formed by different morphological processes.

	Morfessor		NMS	
Category	Prec.	Rec.	Prec.	Rec.
monomorphemic	0.66	0.66	0.79	0.79
reduplication	0.58	0.51	0.85	0.86
affixation	0.92	0.90	0.70	0.70
compounding	0.88	0.91	0.84	0.84

each potential boundary site independently. Precision in this context refers to the proportion of the sites identified by the learner as containing a boundary that also contain a boundary in the gold standard segmentation. Recall refers to the proportion of the sites containing a boundary in the gold standard segmentation that are identified by the learner as containing a boundary. We take a macro-averaging approach: we calculate precision and recall separately for each word, then average each metric across all words in each category. If precision and recall are both undefined for a word (i.e., if the gold standard segmentation contains no boundaries and the learner does not identify any), we set them both to 1; if only one metric is undefined, we set that metric to 0.

3.2.2 Results

The macro-averaged precision and recall for Morfessor and NMS across the four categories of words are shown in Table 1.

For monomorphemic words, both learners show indications of oversegmentation, via low precision and recall that result from placing boundaries where they shouldn't exist. NMS appear to show less oversegmentation than Morfessor, suggesting that they may be more sensitive to word minimality constraints based on moraic weight (Bauer, 1993, p. 544). This tendency toward oversegmentation does not stand out for either learner across other categories: precision and recall are fairly balanced for both learners across all categories, indicating a general balance between oversegmentation and undersegmentation.

For words formed by reduplication, a templatic process, NMS show better performance than Morfessor. This difference is made even clearer when considering performance on reduplication in relation to affixation and compounding (concatenative processes): for Morfessor, performance on reduplication is notably worse than performance on affixation and compounding, but for NMS, it is not. This result suggests that NMS may be sensitive to abstract reduplication templates that Morfessor cannot capture (Todd et al., 2022), and thus that their recognition of such templates may boost implicit learning above and beyond that expected from simple statistical learning of recurrent forms. In support of this suggestion, we found that Morfessor has worst performance on the subset of words formed by total reduplication, the most salient reduplication template, whereas NMS has best performance on this subset (precision/recall for Morfessor: 0.35/0.36; for NMS: 0.95/0.97).

For words formed by affixation and compounding, both concatenative processes, Morfessor performs well, suggesting that such words facilitate implicit learning of morphs via naive statistical learning. Nevertheless, it is somewhat surprising that Morfessor did not perform even better for these words, given that they exactly match the assumptions of its underlying generative model. This suggests that the morphological structure of Māori, as captured by the gold standard segmentations, may be cued by more than just the statistical recurrence of forms (Todd et al., 2019; Panther et al., 2024); we return to this point in Analysis 2 (Section 4).

NMS perform slightly worse than Morfessor on words formed by compounding, and notably worse on words formed by affixation. One possible interpretation of this result is that NMS are not as good at tracking statistical recurrence as Morfessor – hence the worse performance on both categories – but make up for this shortcoming to some extent in compounds by being sensitive to additional cues to morphological structure (Panther et al., 2024). The fact that NMS' difficulties are concentrated in words formed by affixation suggests that they may struggle specifically with recognizing affixes as independent of stems. A finer-grained inspection suggests that this may be related to issues of affix position, allomorphy, and/or frequency: NMS perform as well as Morfessor on words containing the highly frequent causative prefix *whaka-* (precision/recall for Morfessor: 0.95/0.93; for NMS: 0.95/0.93), which has no allomorphs, but perform worse on words containing passive or nominalizing suffixes (precision/recall for Morfessor: 0.90/0.89; for NMS: 0.59/0.59), which have many allomorphs, including some that are quite infrequent.

3.3 Analysis 1B: Affix recovery

To dig further into potential sources of issues with segmenting words formed by affixation, we analyze the ability of Morfessor and NMS to recover different affixes by segmenting them off. This analysis separates the causative prefix from passive and nominalizing suffixes, and subdivides passive and nominalizing allomorphs into smaller groups.

3.3.1 Methods

The affixes we analyze are organized into groups based on word position, status as default/non-default allomorph, and consistency with a major phonological template. The groups also vary in frequency. We define the type frequency of an affix group as the proportion of the 19,595 words for which Oh et al.'s (2020) MS segmented off a morph with the same form as some affix in the group, at the appropriate word edge. We similarly define the token frequency of an affix group as the proportion of tokens in the MAONZE corpus (King et al., 2011) and the Māori Broadcast Corpus (Boyce, 2006) that correspond to words for which the MS separated off some affix from the group.[8] Type frequency is relevant for Morfessor, and both type and token frequency may be relevant for NMS.

For each affix group, we measure the rate at which Morfessor and NMS successfully recover affixes in that group by segmenting them off words. We assign each word in the affixation category to one or more groups based on the affix(es) in its gold standard segmentation. For a word in a given group, a learner successfully recovers the affix pertaining to that group if their segmentation contains a boundary at the site between the affix and the rest of the word, without also containing any boundaries at sites within the affix. The segmentation of the stem is irrelevant: a learner can successfully recover an affix from a word even if their segmentation of the rest of the word does not match that represented by the gold standard segmentation. We measure the rate of affix recovery for a group as the proportion of words in the group for which the affix is successfully recovered.

3.3.2 Results

The affix recovery rates for each learner across the various affix groups are shown in Table 2.

[8]We follow Oh et al. (2020) in using Simple Good-Turing smoothing (Gale and Sampson, 1995) to ensure that words from the dictionary that were not mentioned in the corpora have a non-zero token frequency.

Both Morfessor and NMS have extremely high recovery rates for *whaka-*. This is not surprising, as it is extremely frequent, in terms of both types and tokens. For NMS, it is also highly salient due to its position at the beginning of verbs that often appear utterance-initially as imperatives (e.g., *whakarongo mai!* 'listen!') and its appearance in place names (e.g., Whakatane) and the well known and highly culturally significant word *whakapapa* 'genealogy' (Oh et al., 2023). There is also reason to believe that NMS may be particularly sensitive to prefixes such as *whaka-* because they have been shown to apply a bimoraic template when segmenting the first morph in a word (Panther et al., 2024).

While Morfessor and NMS have near-identical rates for the causative prefix *whaka-*, their recovery rates for allomorphs of the passive and nominalizing suffixes diverge, with NMS being less successful than Morfessor. One possible reason for this divergence is that NMS may be less sensitive to suffixes than prefixes, since the bimoraic template that facilitates sensitivity to prefixes operates from left to right and thus may not consistently align with suffixes. Another possible reason may stem from NMS being sensitive to token frequency rather than just type frequency like Morfessor, as the passive and nominalizing suffixes have much lower token frequencies than type frequencies, both in absolute terms and in relation to the corresponding frequency of *whaka-*. From a statistical learning perspective, a morph needs to be experienced sufficiently often in a range of environments before a learner can reliably identify and extract it, so lower experiential frequency by NMS compared to Morfessor would yield noisier segmentations. This difference is magnified by the fact that Morfessor has perfect memory of all types it has encountered at each point of the learning process, which is not the case for NMS.

Zooming into the allomorphs, for Morfessor there is a clear separation between default and non-default. This separation is driven by frequency: allomorphs in a given affix group can be recovered reliably if and only if they recur across sufficiently many types. After adjusting for frequency, there are no major differences between affix groups based on consistency with a major phonological template, nor phonological shape generally (e.g., passive CVV vs. nominalizing CVCV: recovery rates 0.964 and 0.966, respectively). This is not surprising: Morfessor's naive statistical learning

Table 2: Affix recovery rates for Morfessor and NMS across different affix groups. Affix groups vary in terms of position, status of allomorphs as default/non-default, consistency of allomorphs with major phonological templates, and frequency of occurrence (proportion of types / tokens affixed by that form).

		Frequency		Affix recovery	
Affix(es)	Allomorph group	type	token	Morf.	NMS
whaka-	–	0.142	0.017	0.983	0.976
-tia, *-tanga*	default, template[5]	0.128	0.006	0.995	0.783
-hia, *-ngia*, *-hanga*	default, template[5]	0.064	0.005	0.995	0.688
-a, *-nga*[6]	default, non-template	0.034	0.011	0.907	0.293
-kia, *-mia*, *-ria*, *-whia*, *-kanga*, *-manga*, *-ranga*	non-default, template	0.017	0.003	0.702	0.553
-ia, *-na*, *-ina*, *-hina*, *-kina*, *-whina*, *-anga*	non-default, non-template	0.016	0.002	0.739	0.370

algorithm has no access to phonological templates, and is based primarily on frequency.

For NMS on the allomorphs, there is also a relationship between affix recovery rate and frequency, but it is more gradient, reflecting differences in the experiential frequency and memory of NMS compared to Morfessor. The correlation is not perfect, however. The affix recovery rate is extremely low for the default allomorphs that are not consistent with a template, in spite of their high token frequency. It is also higher than expected for the non-default allomorphs that are consistent with a major phonological template, in comparison to those that have almost identical frequency but are not consistent with a template. These results suggest that NMS are sensitive to major phonological templates, giving them an advantage in recognizing allomorphs that are consistent with them.

Furthermore, since the default allomorphs that are not consistent with a template are also short – with one simply having the shape V – the fact that NMS recover them less successfully suggests a sensitivity to phonological shape generally. That is, NMS may find morphs less salient the less phonological content they have and/or the less their syllables resemble the canonical CV shape. This suggestion is further supported by the fact that NMS are less successful at recovering passive suffixes with a CVV shape than nominalizing suffixes with a CVCV shape (rates 0.669 and 0.866, respectively).

4 Analysis 2: Other cues

Analysis 1 showed that NMS are sensitive to templates, both at the word level (reduplication) and at the morph level (minimality constraints; allomorphs that follow a phonological template or feature syllables with canonical CV shape). Morfessor shows no such sensitivity, as its underlying generative model does not incorporate templates, and thus underperforms when segmenting words that invoke templates in some way.

However, templates appear not to be the only reason that Morfessor underperforms. In Section 3.2.2, we observed that Morfessor's performance on compounds was lower than might be expected, given that they follow its underlying assumption of morphological concatenativity. Based on this observation, we suggested that the morphological structure of Māori may be cued by more than the statistical recurrence of forms, consistent with previous results showing that MS segmentations are sensitive to aspects such as the presence of long vowels (Todd et al., 2019; Panther et al., 2024).

Here, we explore this suggestion further by comparing Morfessor's performance on real Māori words, which may contain such additional cues to morphological structure, to its performance on artificially constructed pseudo-Māori words, which are governed by the same patterns of statistical recurrence of morphs but lack any additional cues to morphological structure. This analysis reveals the extent to which such additional cues exist in real Māori and the extent to which they present issues for Morfessor. In doing so, it generalizes conclusions from Section 3 that the suitability of Morfessor to a particular language – and, by extension, the extent to which statistical learning by non-speakers of that language may be based purely on tracking of statistical recurrence – is dependent upon the morphological structure of the language.

4.1 Data

In this analysis, we focus entirely on words that follow Morfessor's underlying assumption of concatenativity. We do not include words that invoke templates at the word or morph level, since the analysis

in Section 3.2 already established that Morfessor underperforms in the presence of such templates.

The analysis is based on the 'polymoraic' group of Panther et al. (2024), excluding words with morphs containing more than 3 syllables. This includes a total of 1,292 words, comprising 1,199 of the 1,204 compounds that we analyzed in Section 3, as well as an additional 93 words that Oh et al.'s (2020) MS analyzed as simplex.

For the analysis of pseudo-Māori, we generated 1,000 different sets of 1,292 words each through concatenating morphs, based on the statistical properties of the 1,292 real Māori words (see Section 4.2). For each set, the generative process provided us with ground-truth segmentations, which we compare to those provided by a Morfessor model trained over the set. For the analysis of real Māori, we similarly compare the gold standard MS segmentations of the 1,292 words to those provided by a Morfessor model trained on those words (as opposed to the full lexicon from Section 3.1).

4.2 Methods

To generate each set of pseudo-Māori words, we used the same probabilistic process as is assumed by Morfessor's underlying generative model. This process works in a bottom-up fashion across several structural levels, first concatenating phonemes into syllables, then concatenating syllables into morphs, and finally concatenating morphs into words. Types of one level are drawn with replacement from an inventory, according to an inverse power law (Zipfian) probability distribution, and concatenated to form a type of the next level. The types at each level are unique: if a proposed type already exists, a new one is generated instead.

We generated each set of pseudo-Māori words with constraints based on real Māori, in two main ways. First, we constrained the pseudo-Māori words to have the same statistical recurrence properties as real Māori, by using an inventory probability distribution at each level that was inferred from the set of real Māori words (see Appendix A for details). Second, we constrained the types at each level to have the same form properties as real Māori. Specifically: at the phoneme level, we used the same 10 consonants and 10 vowels as real Māori (Section 2.3); at the syllable level, we only generated syllables of shape CV and V; at the morph level, we generated the same number of monosyllabic, disyllabic, and trisyllabic morph types (respectively) as there are in the real Māori set of words; and at the word level, we used each real Māori word as a template for a pseudo-Māori word, ensuring that they matched in terms of the number of morphs and the number of syllables in correspondingly-ordered morphs.

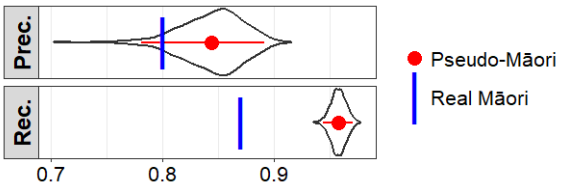

Figure 1: Distributions of macro-averaged precision and recall for Morfessor's segmentations of 1000 sets of pseudo-Māori words, in comparison to its performance on corresponding words from real Māori (blue lines). Red points show mean performance on pseudo-Māori and red lines show 95% percentile intervals.

As in Section 3.2.1, our analysis is based on comparing Morfessor segmentations to a gold standard. We again use macro-averaged boundary precision and recall as the metric for this comparison.

4.3 Results

Figure 1 shows the distributions of macro-averaged precision and recall for Morfessor's segmentations on the 1000 sets of pseudo-Māori words, together with the precision and recall for its segmentations of the corresponding real Māori words (when training is restricted just to those words). It is immediately apparent that recall is higher than precision, indicating occurrences of oversegmentation that are not balanced by undersegmentation as was the case in Section 3.2.2. This is likely a consequence of the training set being much smaller (1,292 words as opposed to 19,595); since the same pattern is seen across real Māori and psuedo-Māori, it does not appear to reflect influences of non-statistical cues to morphological structure.

Morfessor is better able to accurately segment pseudo-Māori than real Māori. Numerically, both precision and recall are higher for pseudo-Māori (mean precision: 0.84; recall: 0.96) than for real Māori (precision: 0.80; recall: 0.87). The advantage for pseudo-Māori is especially strong for recall, where performance on all 1,000 sets of words far exceeds that on real Māori. This strong advantage in recall is not driven by increased oversegmentation of pseudo-Māori relative to real Māori, because it is not accompanied by a concomitant disadvantage in precision; rather, it reflects the fact

that boundaries in pseudo-Māori are cued by recurrence statistics, which Morfessor tracks. That is, Morfessor is best able to segment words when they come from a language that closely adheres to the statistical principles of structure that it assumes.

It follows that Morfessor's worse performance on real Māori is likely due to failure to identify boundaries that are cued by something other than morph recurrence statistics. This result therefore confirms the suggestion from Section 3.2.2 that the morphological structure of Māori may have alternative cues, though it does not indicate precisely what they may be. Past research has shown that NMS are sensitive to cues such as bimoraic templates and the presence of long vowels in the segmentation of compounds (Panther et al., 2024), and it is likely that this sensitivity explains why they were more successful at segmenting compounds than affixed words in Analysis 1A.

5 Discussion & conclusions

We have examined morphological segmentations of Māori by Morfessor and non-Māori-speaking New Zealanders (NMS), across words formed through a variety of morphological processes, to assess the ways in which they are affected by structural factors and the extent to which they have such effects in common. Our results show that both learners are affected by linguistic structure. In some circumstances, they are affected similarly; for example, both are successful in segmenting words formed by concatenative morphological processes (Analysis 1A), especially when highly frequent morphs are involved (Analysis 1B). In other circumstances, they are affected in opposite ways; for example, Morfessor suffers decreased segmentation performance on words that are formed via templatic processes (Analysis 1A) or that cue morphological structure by means other that statistical recurrence of forms (Analysis 2), whereas NMS see increased performance in such cases.

These similarities and differences are important when considering the nature of human statistical learning of morphological segmentation. Since Morfessor's learning is underpinned by a set of well defined assumptions and principles (Section 2.2), the extent to which its performance aligns with that of NMS may be taken to reflect the extent to which NMS' learning is underpinned by those same assumptions and principles. The similarities affirm that NMS undergo statistical learning, identifying and extracting statistically recurrent forms to build a memory-store of morphs. At the same time, the differences show that learning for NMS does not just involve tracking statistical recurrence, but also involves inducing abstract templates about the formation of words and the shapes of (allo)morphs, as well as developing sensitivities to prominent features such as the presence of long vowels (Panther et al., 2024). These findings echo results showing that adults and infants attend to phonological templates when learning to segment artificial languages through incidental exposure (Peña et al., 2002; Marchetto and Bonatti, 2013).

On a practical front, the similarities and differences in the segmentation performances of Morfessor and NMS suggest that human statistical learning of morphological structure can be appropriately modeled by unsupervised machine learning, but perhaps only to a first approximation, depending on the underlying assumptions of the model. When the morphological structures closely follow those assumed by the model, the morphs that the model learns can reflect the cognitive units that humans seem to operate over (e.g., Virpioja et al., 2018; Lehtonen et al., 2019). But when morphological structures vary too widely from those assumed by the model – either within a language, based on words formed by different processes, or across languages – there is the potential for the model to miss factors that are salient to humans but that it is not equipped to handle. This is especially important as different models have different underlying assumptions, which can respond differently to variation in morphological structure (Loukatou et al., 2022).

The differences in the segmentation performances of Morfessor and NMS across words of different morphological structures not only inform the use of unsupervised morphological segmentation models as cognitive models, but also highlight potential factors that could be incorporated into segmentation models to improve their results. For example, inspired by the observation that reduplication templates are salient to humans but not to Morfessor, Todd et al. (2022) show that adding reduplication templates to Morfessor improves its ability to find reduplication in Māori words. Similarly, future research that dissects NMS' underlying learning mechanisms could reveal additional generalizable factors that help improve the cross-linguistic applicability of unsupervised models.

Limitations

While we believe our results to be informative about the effect of language structure on the construction of the NMS proto-lexicon, there are several limitations that could be addressed in future work to clarify and extend them.

First, the gold standard data may not strictly reflect morphological segmentations. One reason for this is that the word-segmentation task through it was obtained taps a form a meta-linguistic knowledge that may not be directly accessible in a consistent manner. However, we do not think this to be a major concern, given that past work using the same task in English (Needle and Pierrehumbert, 2018) found that participants' segmentations matched the underlying morphological structure 88% of the time, and given that we filtered the words used in the analysis to only include those where the gold-standard segmentation is consistent with the assumed morphological structure. We also do not see a better option than eliciting meta-linguistic judgments in this case: the largest group of morphologically complex words in Māori is compounds (Bauer, 1993; Todd et al., 2019), which are not decomposed in any dictionary or large word list of which we are aware.

Second, and relatedly, the gold standard data may contain idiosyncracies, since it was provided by a single MS. While the MS was instructed to split words into parts in a way that they think most Māori speakers would agree with, it is extremely unlikely that their segmentations would all be universally shared. To address this limitation, it would be necessary to repeat the word-segmentation task with many more MS, like we did for NMS.

Third, our comparison of Morfessor and NMS may be complicated by differences between them. For example, Morfessor has perfect memory about all forms of the language and its segmentations of them, but NMS are unlikely to have encountered all words of the language, let alone remember those encounters. Similarly, Morfessor is trained on isolated unique types, whereas NMS experience connected tokens. Morfessor's knowledge is also limited to its Māori training data, whereas NMS also have knowledge of at least one other language (English). It remains to be seen how well Morfessor does when trained on data that resembles what NMS are exposed to, including connected tokens of both Māori and English, and how it may be affected by memory constraints.

Acknowledgements

We thank Forrest Panther for providing the NMS segmentation data. This work used computational facilities purchased with funds from the National Science Foundation (CNS-1725797) and administered by the Center for Scientific Computing (CSC). The CSC is supported by the California Nano Systems Institute and the Materials Research Science and Engineering Center (MRSEC; NSF DMR 2308708) at UC Santa Barbara.

References

Richard N. Aslin. 2017. Statistical learning: A powerful mechanism that operates by mere exposure. *Wiley Interdisciplinary Reviews: Cognitive Science*, 8(1–2):e1373.

Laurie Bauer and Winifred Bauer. 2012. The inflection-derivation divide in Māori and its implications. *Te Reo*, 55:3–24.

Winifred Bauer. 1993. *Maori*. Routledge, London.

Mary Teresa Boyce. 2006. *A Corpus of Modern Spoken Māori*. Unpublished doctoral dissertation, Victoria University of Wellington.

Hermann Bulf, Scott P. Johnson, and Eloisa Valenza. 2011. Visual statistical learning in the newborn infant. *Cognition*, 121(1):127–132.

Mathias Creutz and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, 4(1):3:1–3:34.

Paul de Lacy. 2003. Maximal words and the maori passive. In *Proceedings of AFLA VIII: The eighth meeting of the Austronesian formal linguistics association*, volume 44, pages 20–39. MIT Linguistics Dept.

Ramy Eskander, Owen Rambow, and Tianchun Yang. 2016. Extending the use of adaptor grammars for unsupervised morphological segmentation of unseen languages. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*, pages 900–910.

Michael C. Frank, Joshua B. Tenenbaum, and Edward Gibson. 2013. Learning and long-term retention of large-scale artificial languages. *PLOS ONE*, 8(1):e52500.

William A. Gale and Geoffrey Sampson. 1995. Good-Turing frequency estimation without tears. *Journal of Quantitative Linguistics*, 2(3):217–237.

Judit Gervain, Francesco Macagno, Silvia Cogoi, Marcela Peña, and Jacques Mehler. 2008. The neonate brain detects speech structure. *Proceedings*

of the National Academy of Sciences, 105(37):14222–14227.

Pierre Godard, Laurent Besacier, François Yvon, Martine Adda-decker, Gilles Adda, Hélène Maynard, Annie Rialland, and Inria Grenoble. 2018. Adaptor grammars for the linguist: Word segmentation experiments for very low-resource languages. In *Proceedings of the 15th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 32–42.

Ray Harlow. 2007. *Māori: A Linguistic Introduction*. Cambridge University Press, Cambridge.

Elizabeth K. Johnson. 2016. Constructing a proto-lexicon: An integrative view of infant language development. *Annual Review of Linguistics*, 2(1):391–412.

Mark Johnson and Thomas L. Griffiths. 2007. Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. In *Advances in Neural Information Processing Systems 19*, pages 641–648.

Peter J. Keegan. 1996. *Reduplication in Maori*. Unpublished MA thesis, University of Waikato.

Jeanette King, Margaret Maclagan, Ray Harlow, Peter Keegan, and Catherine Watson. 2011. The MAONZE project: Changing uses of an indigenous language database. *Corpus Linguistics and Linguistic Theory*, 7(1):37–57.

Victor Krupa. 1968. *The Maori Language*. Nauka, Moscow.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. arXiv preprint arXiv:1804.10959.

Minna Lehtonen, Matti Varjokallio, Henna Kivikari, Annika Hultén, Sami Virpioja, Tero Hakala, Mikko Kurimo, Krista Lagus, and Riitta Salmelin. 2019. Statistical models of morphology predict eye-tracking measures during visual word recognition. *Memory & Cognition*, 47:1245–1269.

Georgia Loukatou, Sabine Stoll, Damian Blasi, and Alejandrina Cristia. 2022. Does morphological complexity affect word segmentation? Evidence from computational modeling. *Cognition*, 220:104960.

Erika Marchetto and Luca L. Bonatti. 2013. Words and possible words in early language acquisition. *Cognitive Psychology*, 67(3):130–150.

John C. Moorfield. 2011. *Te Aka: Māori-English, English-Māori Dictionary*, 3rd edition. Pearson, Auckland.

Jeremy Needle and Janet B. Pierrehumbert. 2018. Gendered associations of English morphology. *Laboratory Phonology*, 9(1):14.

Céline Ngon, Andrew Martin, Emmanuel Dupoux, Dominique Cabrol, Michel Dutat, and Sharon Peperkamp. 2013. (Non)words, (non)words, (non)words: Evidence for a protolexicon during the first year of life. *Developmental Science*, 16(1):24–34.

Yoon Mi Oh, Simon Todd, Clay Beckner, Jennifer Hay, and Jeanette King. 2020. Non-Māori-speaking New Zealanders have a Māori proto-lexicon. *Scientific Reports*, 10(1):22318.

Yoon Mi Oh, Simon Todd, Clay Beckner, Jennifer Hay, and Jeanette King. 2023. Assessing the size of non-Māori-speakers' active Māori lexicon. *PLoS ONE*, 18(8):e0289669.

Forrest Panther, Wakayo Mattingley, Jennifer Hay, Simon Todd, Jeanette King, and Peter J. Keegan. 2024. Morphological segmentations of non-Māori speaking New Zealanders match proficient speakers. *Bilingualism: Language and Cognition*, 27(1):1–15.

Forrest Panther, Wakayo Mattingley, Simon Todd, Jennifer Hay, and Jeanette King. 2023. Proto-lexicon size and phonotactic knowledge are linked in non-Māori speaking New Zealand adults. *Laboratory Phonology*, 14(1).

ʻŌiwi Parker Jones. 2008. Phonotactic probability and the māori passive. In *Proceedings of the Tenth Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pages 39–48.

Marcela Peña, Luca L. Bonatti, Marina Nespor, and Jacques Mehler. 2002. Signal-driven computations in speech processing. *Science*, 298(5593):604–607.

Bruna Pelucchi, Jessica F. Hay, and Jenny R. Saffran. 2009. Statistical learning in a natural language by 8-month-old infants. *Child Development*, 80(3):674–685.

Patrick Rebuschat. 2015. *Implicit and Explicit Learning of Languages*. John Benjamins Publishing Company.

Jorma Rissanen. 1978. Modeling by shortest data description. *Automatica*, 14(5):465–471.

Aku Rouhe, Stig-Arne Grönroos, Sami Virpioja, Mathias Creutz, and Mikko Kurimo. 2022. Morfessor-enriched features and multilingual training for canonical morphological segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 144–151.

Jenny R. Saffran. 2001. Words in a sea of sounds: the output of infant statistical learning. *Cognition*, 81(2):149–169.

Jenny R. Saffran, Richard N. Aslin, and Elissa L. Newport. 1996. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909.

Tuomas Teinonen, Vineta Fellman, Risto Näätänen, Paavo Alku, and Minna Huotilainen. 2009. Statistical language learning in neonates revealed by event-related brain potentials. *BMC Neuroscience*, 10(1):21.

Simon Todd, Chadi Ben Youssef, and Alonso Vásquez-Aguilar. 2023. Language structure, attitudes, and learning from ambient exposure: Lexical and phonotactic knowledge of Spanish among non-Spanish-speaking Californians and Texans. *PLOS ONE*, 18(4):e0284919.

Simon Todd, Annie Huang, Jeremy Needle, Jennifer Hay, and Jeanette King. 2022. Unsupervised morphological segmentation in a language with reduplication. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 12–22.

Simon Todd, Jeremy Needle, Jeanette King, and Jennifer Hay. 2019. Quantitative insights into Māori word structure. Paper presented at the Annual Meeting of the Linguistic Society of New Zealand.

Sami Virpioja, Minna Lehtonen, Annika Hultén, Henna Kivikari, Riitta Salmelin, and Krista Lagus. 2018. Using statistical models of morphology in the search for optimal units of representation in the human mental lexicon. *Cognitive Science*, 42(3):939–973.

Sami Virpioja, Peter Smit, Stig-Arne Grönroos, and Mikko Kurimo. 2013. Morfessor 2.0: Python implementation and extensions for Morfessor Baseline. Technical report, Department of Signal Processing and Acoustics, Aalto University, Helsinki.

Sami Virpioja, Ville T. Turunen, Sebastian Spiegler, Oskar Kohonen, and Mikko Kurimo. 2011. Empirical comparison of evaluation methods for unsupervised learning of morphology. *Traitement Automatique des Langues*, 52(2):45–90.

John N. Williams. 2020. The neuroscience of implicit learning. *Language Learning*, 70:255–307.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.

A Generating pseudo-Māori: Details

This appendix describes the process through which we inferred statistical recurrence properties of Māori, to use in the generation of pseudo-Māori.

We derived inventories at each level – unique phonemes, syllables, morphs, and words – from the segmentations provided by Oh et al.'s (2020) fluent Māori speaker. To get the frequency distribution over types at one level, we counted occurrences within unique types at the next level. That is, we counted the number of unique syllables that each phoneme occurred in; the number of unique morphs that each syllable occurred in; and the number of unique words that each morph occurred in. We sorted each distribution by count, to obtain rank and frequency for each type, and fit an inverse power law $f(x) = ab^{-x}$ to predict frequency from rank, using nonlinear least squares.

To sample in the generative process, we sorted the types in random order and treated those orders as ranks, overlaying the frequency from the inverse power law and then normalizing to obtain a probability distribution.

Different Tokenization Schemes Lead to Comparable Performance in Spanish Number Agreement

Catherine Arnett*[1], Pamela D. Rivière*[2], Tyler A. Chang[2,3], Sean Trott[2]
[1]Department of Linguistics,
[2]Department of Cognitive Science,
[3]Halıcıoğlu Data Science Institute
UC San Diego
{ccarnett, pdrivier, tachang, sttrott}@ucsd.edu

Abstract

The relationship between language model tokenization and performance is an open area of research. Here, we investigate how different tokenization schemes impact number agreement in Spanish plurals. We find that morphologically-aligned tokenization performs similarly to other tokenization schemes, even when induced artificially for words that would not be tokenized that way during training. We then present exploratory analyses demonstrating that language model embeddings for different plural tokenizations have similar distributions along the embedding space axis that maximally distinguishes singular and plural nouns. Our results suggest that morphologically-aligned tokenization is a viable tokenization approach, and existing models already generalize some morphological patterns to new items. However, our results indicate that morphological tokenization is not strictly required for performance.

1 Introduction

In natural language processing (NLP) pipelines, **tokenizers** segment unstructured text into smaller, discrete constituents ("tokens") for further processing. Importantly, different tokenizers can incur performance and efficiency trade-offs. Assigning a unique token to each word in a corpus may lead to high-precision semantic representations, but the resulting models might be less robust to unseen words and require more computational resources.

Most existing tokenizers allow words to be decomposed into subword tokens (Sennrich et al., 2016; Kudo and Richardson, 2018). They can do so along morphological boundaries (e.g. *books* to ['book', '##s']), but this behavior is not guaranteed. Segmenting words into their lemmas and morphemes might simultaneously allow models to more robustly learn morphosyntactic patterns, more efficiently represent such patterns, and better generalize to novel words. (An analogous question concerning the storage of whole words vs. learning generalizable rules exists within human psycholinguistics research, e.g., Ullman, 2016).

In the current work, we ask whether and how the tokenization strategy employed facilitates successful language model predictions. We evaluate the effect of three types of plural noun tokenization in Spanish—single-token plurals, morphemically-tokenized plurals, and non-morphemically-tokenized plurals—in the context of a masked article prediction task (§4).[1] We focus on tokenization schemes for plural forms in Spanish, as it offers relatively simple and frequent examples of morphologically complex words. Spanish leverages two primary plural marking strategies, which are highly predictable for any given lemma. We specifically focus on cases where the plural form is composed of the singular form with the addition of '-s' or '-es'.

We find that tokenization schemes are differentially successful, although the effect is small, and article agreement accuracy is high across all tokenization types. Artificial tokenization schemes, where we coerce an initially single-token or non-morphemically-tokenized plural into a morphemic representation, leads to successful task performance, but does not improve performance beyond the original tokenization scheme. In an exploratory analysis, we compare singular and plural form embeddings across all tokenization schemes. We find axes with high overlap between all plural forms (regardless of tokenization scheme) and high discriminability between plural and singular forms, but other axes can still separate different plural tokenization schemes. This work contributes to a growing literature examining the impact of tokenization on the language

*Equal contribution.

[1]Note that this categorization scheme mirrors an approach taken in contemporaneous work, using the labels "vocab", "morph", and "alien", respectively.

Proceedings of the The 21st SIGMORPHON workshop on Computational Morphology, Phonology, and Phonetics, pages 32–38

modeling objective. Code and data are available: https://github.com/catherinearnett/spanish-plural-agreement.

2 Related Work

Several studies have investigated morpho-syntactic agreement in BERT-style models across multiple languages (Linzen et al., 2016; Mueller et al., 2020; Edmiston, 2020; Pérez-Mayos et al., 2021, inter alia), finding generally high agreement accuracy. In a subject-verb agreement task, however, BETO incurs a relatively high rate of agreement errors for certain Spanish nouns (despite the ability to extend number agreement to novel words; Haley, 2020). It is unclear to what extent degraded performance is attributable to tokenization scheme, but the word "comanas"—listed as an example of a frequently mis-numbered word—is tokenized non-morphemically into ['coman', '##as'].

Indeed, recent work has demonstrated that morphologically-aware tokenization improves NLP model performance on a variety of downstream benchmarks (Park et al., 2020; Hofmann et al., 2021; Toraman et al., 2023; Jabbar, 2024; Uzan et al., 2024). Most relevantly, Batsuren et al. (2024) devise a tool to classify English words in terms of whether they are stored as single tokens ("vocab"), as multiple morphemic tokens ("morph"), or as multiple non-morphemic tokens ("alien"). The authors find that how multi-morphemic English words are tokenized is correlated with the language model's downstream performance on several tasks.

Following Batsuren et al. (2021), our work investigates how the tokenization of Spanish nouns affects language model predictions involving a specific morphosyntactic rule, providing insight into how morphologically-aware tokenization affects NLP model performance.

3 Model and Data

All experiments use BETO, a Spanish pre-trained BERT model (Cañete et al., 2020) with 110M parameters trained on approximately 3B words. BETO uses a SentencePiece tokenizer (Kudo and Richardson, 2018) with a 32K vocab size.

3.1 Data

All plural nouns and their singular form lemmas were extracted from the AnCora Treebanks (Alonso and Zeman, 2016). Plurals were categorized according to their affix. Nouns ending in vowels use the plural suffix -s, while nouns ending in consonants use the suffix -es. Plurals were also annotated for their grammatical gender by a native Spanish speaker. Irregular nouns, misspellings, and words not listed in the Real Academia Española (RAE) online dictionary were excluded.

3.2 Identifying Tokenization Type

We created three lists of plurals: one-token (n=1247), multi-token morphemic (n=508), and multi-token non-morphemic (n=627). One-token plurals are stored as single tokens in the tokenizer's vocabulary. We then categorized multi-token plurals as morphemic or non-morphemic. If tokenization followed morpheme boundaries (e.g., *naranjas* as ['naranja', '##s']), the noun was categorized as morphemic; if not, it was categorized as non-morphemic (e.g., *neuronas* is tokenized as ['neuro', '##nas']).

3.3 Relationship of Tokenization to Frequency

Using oral frequency measures for 2071 target plural wordforms available in a corpus of over 3M spoken words (Alonso et al., 2011), we examined the relationship between a wordform's frequency and how it was tokenized. A linear model predicting Log Frequency from Tokenization Scheme explained significant variance [$R^2 = 0.33$]. With MORPHEMIC level as a reference class (i.e., intercept), the NON-MORPHEMIC plural nouns were significantly less frequent [$\beta = -0.18, SE = 0.03, p < .001$], while the SINGLE-TOKEN plural nouns were significantly more frequent [$\beta = 0.59, SE = 0.03, p < .001$]. As expected, the frequency of a wordform was likely a major factor in how it was tokenized (see also Appendix A.2).

Due to the relationship between tokenization scheme and wordform frequency, we carried out several supplementary analyses to determine the extent to which frequency was a confound in the results presented in Section 4. We found two key results: first, BETO's predictions were indeed more accurate for more frequent wordforms; second, however, BETO's predictions were still more accurate for some of the original tokenization schemes than others, even controlling for wordform frequency (see Appendix A.2 for details).

3.4 Artificial Tokenization Procedure

To investigate the effect of tokenizing a wordform at the morpheme boundary, we artificially tokenized single-token and multi-token non-

morphemic plural nouns by concatenating the token for the appropriate affix (e.g., "##es") onto the token(s) for the singular noun (Table 1).

Morpheme Boundary	Original Tokenization	Artificial Tokenization
mujer+es	['mujeres']	['mujer', '##es']
patrono+s	['patr, '##onos']	['patr, '##ono', '##s']

Table 1: Artificial tokenizations for the words *mujeres* 'women' (*mujer*), and *patronos* 'employers' (*patrono*).

4 Study: Article-Noun Agreement

Our primary research question concerned the impact of the original tokenization (TOKENIZATION SCHEME) on an article agreement task, similar to that implemented by Linzen et al. (2016). In Spanish, articles must agree with the *number* of the noun (e.g., *la mujer* vs. *las mujeres*); learned representations for the target noun should thus be conducive to predicting article number. We asked:

1. How does the initial tokenization scheme of a plural noun impact the language model's ability to predict the correct article?

2. Does our *artificial* tokenization scheme provide sufficient information to facilitate successful agreement?

3. How does the success of our artificial tokenization scheme compare to the original tokenization scheme for those nouns?

4.1 Method

Implementational details for the masked article prediction task are available in Appendix A.1. Agreement was assessed by taking the logarithm of the relative probability of a plural vs. singular article as predicted by a given noun. For a given wordform (e.g., *mujeres*), a positive log-odds indicated a higher probability was assigned to the plural article, while a negative log-odds indicated a higher probability was assigned to the singular article. A *singular* noun should be associated with a more negative log-odds, while a *plural* noun should be associated with a more positive log-odds. We considered both DEFINITE and INDEFINITE articles (ARTICLE TYPE) for each wordform; the log-odds calculation was performed separately for each type.

Accounting for the different presentations of each wordform (i.e., definite vs. indefinite article; original vs. artificial tokenization), our final dataset had 13,276 observations in total, each with an accompanying *log-odds* ratio. All data and visualizations were analyzed in R; mixed effects models were fit using the *lme4* package (Douglas Bates et al., 2015). Maximal random effects structures were fit where possible, and reduced as needed for model convergence.

4.2 Results

4.2.1 Impact of Initial Tokenization

We first asked whether the original tokenization scheme used for plural nouns affected successful agreement. We fit a mixed model with Log Odds as a dependent variable, fixed effects of Tokenization Scheme and Word Number (and an interaction between the two), fixed effects of Article Type, and random intercepts for each word lemma and sentence. This model explained significantly more variance than a model omitting only the interaction $[\chi^2(2) = 6.54, p = .04]$, suggesting that different tokenization schemes were differentially successful in predicting the appropriate article. Note that this interaction was independent from the effect of wordform frequency (see Appendix A.2).

However, as depicted in Figure 1, this effect was quite small. Accuracy was near ceiling for all tokenization types, i.e., the Log Odds was larger than 0 for plural nouns and smaller than 0 for singular nouns (see also Table 2). Thus, our results do not suggest that morphologically-aligned tokenization is required for good agreement performance.

Original Tokenization	Original	Artificial
Morphemic	0.97	—
Non-morphemic	0.98	0.96
Single-Token	0.98	0.97

Table 2: Accuracy scores for *plural nouns* only, using either the original tokenization scheme for that class of nouns or the artificially-induced morphemic scheme.

4.2.2 Success of Artificial Tokenization

Next, we artificially tokenized plural nouns that would otherwise be tokenized non-morphemically or as a single-token. To quantify the success of this procedure, we fitted a linear mixed-effects model predicting Log Odds with fixed effects of Article Type, Word Number, Tokenization Scheme, and Affix ('##s' or "##es"), as well as random intercepts for word lemma and sentence.

This model explained significantly more variance than a model omitting only Word Number

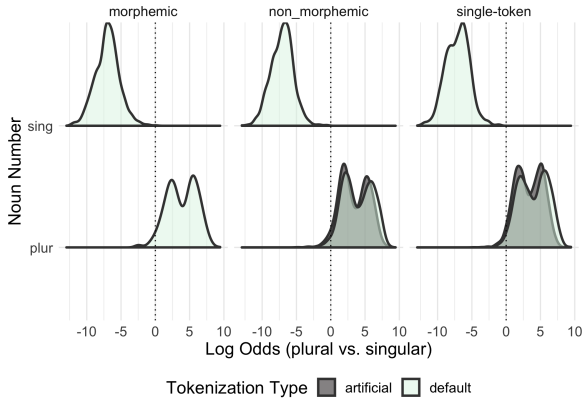

Figure 1: Log-odds varied significantly as a function of noun number (*singular* vs. *plural*). The extent of this variance interacted (weakly) with initial tokenization (*morphemic* vs. *non-morphemic* vs. *single-token*) and with whether the *original* or *artificial* tokenization procedure was used. Larger log-odds indicate higher probabilities of the plural article.

$[\chi^2(1) = 11988, p < .001]$, indicating that the artificial tokenization procedure still led to good article number agreement performance: Log Odds were significantly different for singular nouns and artificially-tokenized plural nouns (see also Figure 1 and Table 2).

4.2.3 Comparing Default vs. Artificial Tokenization Schemes

Finally, restricting our analysis to plural forms, we asked whether a higher Log Odds was assigned to *artificially tokenized* plural nouns than ones using the default scheme. We fitted a linear mixed-effects model with fixed effects of Tokenization Scheme (artificial or original), Affix, and Original Tokenization Scheme (as well as random intercepts for word lemma, sentence, and wordform, and by-lemma random slopes for Tokenization Scheme). This model did explain more variance than a model omitting only Tokenization Scheme $[\chi^2(1) = 141.81, p < .001]$. Critically, however, the Log Odds for the artificially tokenized plural nouns was *lower* ($M = 3.38, SD = 2$) than when using the default tokenization ($M = 3.95, SD = 2.15$). In other words, the artificially-induced morphemic tokenization was successful, but less so than relying on the original scheme for those nouns.

5 Linear Discriminant Analysis (LDA)

To identify potential causes for the observed agreement patterns across noun types (singular vs. different plural tokenizations), we considered the embeddings of those nouns in the language model representation space. We took each noun's mean embedding across the last four (out of twelve) BETO Transformer layers, averaging over all tokens in the noun. To minimize confounds from averaging embeddings over different numbers of tokens, we considered only two-token plurals in all multi-token scenarios for embedding analyses.

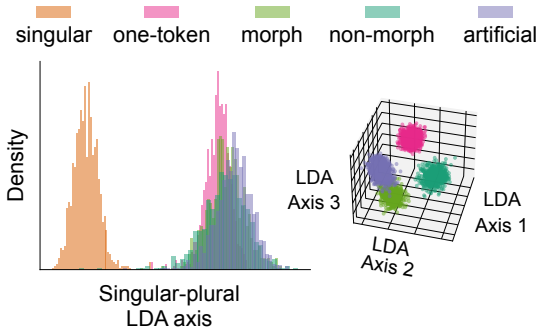

Figure 2: LDA for singular and plural embeddings reveals axes of overlap (*left*) and discriminability (*right*) for differentially tokenized plural forms.

We first identified the linear axis that maximally separated single-token singular from plural nouns. To do this, we ran linear discriminant analysis (LDA) with two classes of embeddings: singular nouns (all single-token) and single-token plural nouns.[2] We then projected all noun representations linearly onto this axis, essentially projecting each embedding into a single value. As expected, we found that singular nouns clustered separately from plural nouns (Figure 2, *left*). Notably, all types of plurals (single-token, artificially tokenized, two-token morphemic, and two-token non-morphemic) patterned together and were not linearly discriminable along this axis. This suggests that the model could rely on similar number agreement mechanisms for different types of plurals, but future work would need to demonstrate causal impacts of this singular-plural axis on number agreement predictions (e.g. as in Mueller et al., 2022).

While the singular-plural LDA axis mapped different plural types to similar values, other axes could separate embeddings for the different plural types. We used LDA to identify the three linear axes that maximally separated the four types of plurals. As shown in Figure 2 (*right*), single-token plurals and two-token non-morphemic plu-

[2]Given n sets of representations, LDA computes $n - 1$ directions in the language model representation space that maximize separation between the sets.

rals were separable from one another and from all other plural types. The artificial and default morphemic plurals had distinct clusters, but they were not entirely separable from one another. This indicates that even though the artificial tokenization was never seen by the model during training, the representations were still quite similar (e.g. due to the presence of the '##s' or '##es' token). The slight separation between these clusters may be driven either by frequency effects or by veridical differences in how the models represent number in the two plural types.

6 Discussion and Conclusion

We assessed whether distinct tokenization schemes impacted the ability of BETO (a Spanish language model) to predict appropriate articles for Spanish plural nouns. Single-token representations facilitated slightly better predictions overall. However, the model did show evidence of generalization consistent with having learned morpheme-like "rules": artificially re-tokenizing plural nouns along morpheme boundaries produced representations amenable to article prediction—despite the language model never having previously observed that sequence of tokens (see Figure 1)—though this approach was slightly less accurate than relying on the original tokenization scheme. This provides further insight into work on language models generalizing morphological patterns (Haley, 2020); however, this does not work equally well for all languages or models (Weissweiler et al., 2023).

Notably, the similar agreement performance across single-token, morphological, non-morphological, and artificially-tokenized plurals could indicate multiple different agreement mechanisms in the model. At least on this task, tokenization along morpheme boundaries was not correlated with improved agreement performance; this is in contrast to other work suggesting that morphologically aware tokenization improves performance, e.g., in machine translation (Macháček et al., 2018) or similarity judgments (Batsuren et al., 2024). Future work might apply causal interventions on different embedding axes (as found in §5), to determine the extent to which the same model subnetworks are involved in number agreement for different types of plural tokenizations, shedding light on the impacts of tokenization on language model processing.

7 Limitations

A key limitation of the current work is scope. Future work could consider additional morphological phenomena, additional languages, and a larger range of language models or tokenization schemes. A second limitation is that the language model's performance was near-ceiling for each category considered. It is possible that different tokenization strategies do in fact impact agreement performance under more challenging conditions, but that the near-ceiling performance on this task made it difficult to detect those differences. Future work could work to develop more challenging tasks for which the model is not at ceiling (as in Linzen et al., 2016), or for which variance in how multimorphemic words are parsed might be expected to contribute more to downstream performance (Batsuren et al., 2024). Finally, our work does not demonstrate the extent to which different tokenizations rely on the same internal mechanisms for agreement in the model (§6), which is a valuable direction for future work.

8 Acknowledgements

Tyler Chang is partially supported by the UCSD Halıcıoğlu Data Science Institute graduate fellowship, and Pamela D. Rivière is supported by UCSD's Chancellor's Postdoctoral Fellowship Program.

References

Héctor Martínez Alonso and Daniel Zeman. 2016. Universal Dependencies for the AnCora treebanks. *Procesamiento del Lenguaje Natural*, 57.

María Angeles Alonso, Angel Fernandez, and Emiliano Díez. 2011. Oral frequency norms for 67,979 Spanish words. *Behavior Research Methods*, 43:449–458.

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2021. MorphyNet: a large multilingual database of derivational and inflectional morphology. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 39–48, Online. Association for Computational Linguistics.

Khuyagbaatar Batsuren, Ekaterina Vylomova, Verna Dankers, Tsetsuukhei Delgerbaatar, Omri Uzan, Yuval Pinter, and Gábor Bella. 2024. Evaluating subword tokenization: Alien subword composition and oov generalization challenge. *arXiv preprint arXiv:2404.13292*.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained BERT model and evaluation data. In *PML4DC at ICLR 2020*.

MM Douglas Bates, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Daniel Edmiston. 2020. A systematic analysis of morphological content in BERT models for multiple languages. *arXiv preprint arXiv:2004.03032*.

Coleman Haley. 2020. This is a BERT. Now there are several of them. Can they generalize to novel words? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 333–341.

Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2021. Superbizarre is not superb: Derivational morphology improves BERT's interpretation of complex words. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3594–3608, Online. Association for Computational Linguistics.

Haris Jabbar. 2024. MorphPiece: A linguistic tokenizer for large language models. *arXiv*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Dominik Macháček, Jonáš Vidra, and Ondřej Bojar. 2018. Morphological and language-agnostic word segmentation for nmt. In *International Conference on Text, Speech, and Dialogue*, pages 277–284. Springer.

Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. 2020. Cross-linguistic syntactic evaluation of word prediction models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5523–5539, Online. Association for Computational Linguistics.

Aaron Mueller, Yu Xia, and Tal Linzen. 2022. Causal analysis of syntactic agreement neurons in multilingual language models. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 95–109, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Kyubyong Park, Joohong Lee, Seongbo Jang, and Dawoon Jung. 2020. An empirical study of tokenization strategies for various Korean NLP tasks. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 133–142, Suzhou, China. Association for Computational Linguistics.

Laura Pérez-Mayos, Alba Táboas García, Simon Mille, and Leo Wanner. 2021. Assessing the syntactic capabilities of transformer-based multilingual language models. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3799–3812, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Cagri Toraman, Eyup Halit Yilmaz, Furkan Şahinuç, and Oguzhan Ozcelik. 2023. Impact of tokenization on language models: An analysis for Turkish. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4):1–21.

Michael T Ullman. 2016. The declarative/procedural model: A neurobiological model of language learning, knowledge, and use. In *Neurobiology of language*, pages 953–968. Elsevier.

Omri Uzan, Craig W Schmidt, Chris Tanner, and Yuval Pinter. 2024. Greed is all you need: An evaluation of tokenizer inference methods. *arXiv preprint arXiv:2403.01289*.

Leonie Weissweiler, Valentin Hofmann, Anjali Kantharuban, Anna Cai, Ritam Dutt, Amey Hengle, Anubha Kabra, Atharva Kulkarni, Abhishek Vijayakumar, Haofei Yu, Hinrich Schuetze, Kemal Oflazer, and David Mortensen. 2023. Counting the bugs in ChatGPT's wugs: A multilingual investigation into the morphological capabilities of a large language model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6508–6524, Singapore. Association for Computational Linguistics.

A Appendix

A.1 Article Agreement Task: Model Inputs

Within the sequence of model inputs, only the article token was masked, and special tokens ([CLS], [SEP]) were included, as in the examples below:

- Example model inputs for original single-tokenizations: "[CLS] [MASK] mujeres [SEP]"

- Example model inputs for artificial (morphemic) tokenizations: "[CLS] [MASK] mujer ##es [SEP]"
- Example model inputs for original non-morphemic multi-tokenizations: "[CLS] [MASK] patr ##onos [SEP]"
- Example model inputs for artificial (morphemic) tokenizations: "[CLS] [MASK] patr ##ono ##s [SEP]"

For each sequence of inputs independently, we obtain BETO's output logits over the target token corresponding to the (1) definite singular, (2) indefinite singular, (3) definite plural, and (4) indefinite plural articles. We subsequently apply softmax normalization to each token's logits to obtain the log probabilities of filling the masked item with a particular article.

A.2 Supplementary Analysis with Log Frequency

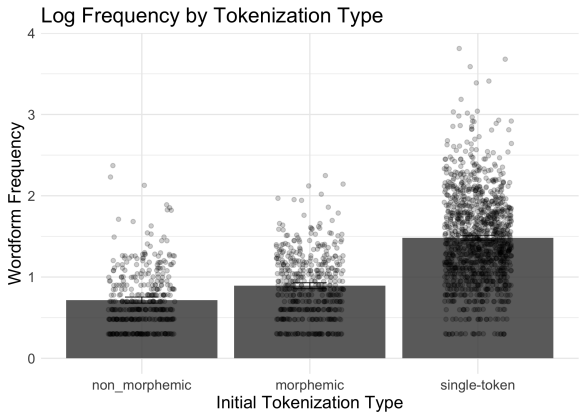

Figure 3: Single-token plurals were significantly more frequent than those tokenized according to morphemic boundaries, which were more frequent than those tokenized according to non-morphemic substrings.

We ran a follow-up analysis asking whether the Log Frequency of a wordform was predictive of agreement success. This analysis had two key goals. First, because Log Frequency was correlated with Tokenization Scheme, we aimed to determine whether the effect of Tokenization Scheme on agreement success was in fact due to effects of token frequency. Second, we were independently interested in whether the language model made better predictions for more frequent wordforms.

We fitted a linear mixed-effects model including fixed effects of Tokenization Scheme, Word Number, and Log Frequency, as well as interactions between Word Number and Tokenization Scheme and between Word Number and Log Frequency. We also included random intercepts for word lemma and sentence. This model explained significantly more variance than a model omitting only the interaction between Log Frequency and Word Number $[\chi^2(1) = 17.89, p < .001]$. The interaction was negative $[\beta = -0.35, SE = 0.08, p < .001]$, i.e., the plural article log-odds were *more* negative for more frequent singular nouns. In other words, the language model made better predictions for more frequent nouns than less frequent nouns.

The full model also explained more variance than a model omitting the interaction between Word Number and Tokenization Scheme $[\chi^2(2) = 11.24, p = .004]$. This indicates that even controlling for wordform frequency, there was an independent effect of how the wordform was initially tokenized on the success of the language model's article predictions.

Ye Olde French: Effect of Old and Middle French on SIGMORPHON-UniMorph Shared Task Data

William Kezerian, Lam An Wyner, Sandro Ansari and **Kristine M. Yu**
Department of Linguistics
University of Massachusetts Amherst
{wkezerian, lamanwyner, alexanderans, kmyu} @umass.edu

Abstract

We offer one explanation for the historically low performance of French in the SIGMORPHON-UniMorph shared tasks. We conducted experiments replicating the 2023 task on French with the non-neural and neural baselines, first using the original task splits, and then using splits that excluded Old and Middle French lemmas. We applied a taxonomy on our errors using a framework based on Gorman et al. (2019)'s annotation scheme, finding that a high portion of the French errors produced with the original splits were due to the inclusion of Old French forms, which was resolved with cleaned data.

1 Introduction

The annual SIGMORPHON-UniMorph shared task on morphological (re)-inflection has been a locus for developing language resources, algorithms, and tasks in the computational morphology community since Cotterell et al. (2016). While the details of the task settings have varied over the years, the basic task has been: given training data consisting of triples ⟨lemma, inflection features, inflected form⟩, (e.g., ⟨*désarmer*, 2.SG.SUBJ, *désarmerais*⟩ for French 'disarm',) train a model to infer inflected forms given pairs ⟨lemma, inflection features⟩.

Already in 2016, Cotterell et al. (2016) noted the remarkable average accuracy achieved in the basic task (95.56% averaged across languages in the top system) and the "surprising" huge performance gap between neural and non-neural approaches (e.g., with the best performing neural approach exceeding the best non-neural one by as much as 60% in accuracy within a language). Seven years and three versions of UniMorph later, the most recent SIGMORPHON-UniMorph shared task, Goldman et al. (2023, p. 120) similarly remarks that performance over individual languages was "quite impressive" and that "all neural systems outperformed the non-neural systems" on average across languages.

However, Goldman et al. (2023) also notes the mysteriously poor performance of systems on French in two senses. First, no system achieved higher than 77.7% accuracy on French. The other languages for which models' accuracies peaked at 80% were Navajo, Ancient Greek, Sanskrit, Belarusian, and Sami. Goldman et al. (2023, p. 121) implicitly note the oddness of French (a suffixing, fusional, high-resource language) in this group: "While there is no one characteristic shared between all of these languages, it is worth noting that this list includes the only two extinct languages tested in this task, and the only mostly prefixing language. Perhaps further development of tailored models could help fill this gap."

Second, neural systems did not outperform non-neural systems on French—while the non-neural baseline achieved 77.7% accuracy, the best performing neural system achieved 74.7% accuracy. In fact, the non-neural baseline was the best performing system in English, Danish, and French. Goldman et al. (2023, p. 120) again point out the oddness of French in this group: "Partial explanation may be the small size of the inflection tables in Danish and English that necessitated inclusion of many lemmas in the training set and may facilitated better generalization ability of the non-neural baseline. **Admittedly, this explanation is not valid for French,[1] but this language was proven difficult in previous shared tasks (Cotterell et al., 2017, 2018) and in other works (Silfverberg and Hulden, 2018; Goldman and Tsarfaty, 2021).**"

Why has French been a particularly challenging language for inflection tasks since it was first added to UniMorph in 2017? In this paper, we show that Old and Middle French lemmas/forms have been erroneously included in French UniMorph data in all SIGMORPHON-UniMorph shared tasks involving

[1] Splits were sampled from 500 lemmas for French, but 2000 for Danish and 3000 for English (Goldman et al., 2023, Table 2).

Proceedings of the The 21st SIGMORPHON workshop on Computational Morphology, Phonology, and Phonetics, pages 39–50

French, as well as Silfverberg and Hulden (2018); Goldman and Tsarfaty (2021). We also provide evidence that including these Old and Middle French forms has caused anomalously poor performance via three replication experiments of the 2023 shared task for French—two excluding Old and Middle French lemmas—and an error analysis of the results.

2 Background

To contextualize our claim that Old and Middle French forms have resulted in poor performance, we will first provide background information on Old and Middle French and explain its presence in Wiktionary, as well as a brief history of poor performance on French in past inflection tasks. Hereafter "Old French" will be used as shorthand to encompass both Old French and Middle French.

2.1 Old French

Old French evolved into Middle French in the 14th century, then to modern French in the 17th century. Old French conjugation tables have been extensively documented in the English edition of Wiktionary (the source of data for `fra`, the French UniMorph data file). The only cited source for these tables is *Dictionnaire de l'ancienne langue française et de tous ses dialectes du IXe au XVe siècle* (Godefroy, 1881), which outlines all of the possible conjugations for Old French verbs.

2.1.1 Old French lemmas and suffixes

Most suffixes used in Old French verb conjugations are not licit verb inflection suffixes in modern French. These include *-ois*, *-oit*, and *oient* in the past imperfect and *â* in past perfect suffixes such as *-astes* and *-asmes*. For more examples, see Table A2 in the Appendix.

Although it is fairly easy to identify Old French verb suffixes, there are no universal patterns that make it clear whether a lemma itself is Old French or modern French. This task requires French linguistic knowledge and investigation into the documentation on the verb.

2.2 Poor performance on French verb inflection in SIGMORPHON-UniMorph and related inflection tasks

SIGMORPHON-UniMorph shared tasks French verbal paradigms were first included in the SIGMORPHON-UniMorph shared inflection task in 2017.[2] In that task (subtask 1), the best-performing system (UE-LMU, neural) scored 89.50% by-form accuracy on French in the high resource setting, cf. 81.50% from the non-neural baseline (Cotterell et al., 2017, Table 12). Among the 52 languages in the task, only 4 had comparably poor performance (Cotterell et al., 2017, Table 9).

In the 2018 SIGMORPHON shared task (task 1), French appeared as a surprise language. The highest accuracy on French in the high resource setting was 90.40% (uzh-2, neural), cf. 82.80% from the non-neural baseline. Only 8 out of 103 languages had comparable or lower performance. (Cotterell et al., 2018, Tables 9, 10, 14). French was also included as part of the French-Occitan pair in the 2019 SIGMORPHON-UniMorph shared Task 1 involving training on high-resource languages to infer inflection on genetically related low-resource languages, but inferring something about French from performance is difficult since performance varied highly by how closely the two languages in the pair were related. After 2019, French was not included in SIGMORPHON-UniMorph shared tasks again until 2023.

While Romanian, Hungarian, and Latin yielded poorer performance than French in both the 2017 and 2018 shared tasks, Gorman et al. (2019, p. 143; Table 4)'s error analysis of the 2017 shared task discovered that all three of these languages suffered from a preponderance of **extraction errors** in how UniMorph parsed Wiktionary's inflectional paradigms that would have impacted performance in both 2017 and 2018. Gorman et al. (2019) did not perform an error analysis of French.

Minimal supervision inflection tasks Goldman et al. (2023, p. 120) also pointed to poor performance on French in Silfverberg and Hulden (2018) and Goldman and Tsarfaty (2021). However, Silfverberg and Hulden (2018) did not report *uniformly* low performance for French verbs across tasks. They trained an encoder-decoder model on 1 to 3 forms randomly sampled from: (i) 1000 randomly sampled inflection tables from UniMorph, or (ii) 1,131 inflection tables from UniMorph that contained items among the 10,000 most frequent word tokens from Al-Rfou' et al. (2013)'s dump of the French edition of Wikipedia. The task was then to generate the remaining missing forms in each

[2]While UniMorph 4.0 (Batsuren et al., 2022) added adjectives and nouns in `fra.segmentation`, all inflection tasks for French discussed in this paper have been only for verbs.

inflection table. When the inflection tables were randomly sampled, accuracy on missing forms in French verbs was the lowest of all 8 languages/part of speech data sets for 1, 2, and 3 forms, e.g., 83.64% for 3 forms, cf. 74.07% for the baseline model, a new implementation of Malouf (2017)'s LSTM model.

But when the inflection tables for training were sampled to contain *the most frequent* forms, accuracy for French verbs was 31.34% for French verbs (cf. 14.34% for the baseline)——in the middle of the pack among the 8 data sets, and higher than for Spanish verbs or Finnish verbs. Moreover, Silfverberg and Hulden (2018) reported one instance of near perfect accuracy for French: 99.50% accuracy in validating their implementation of Malouf (2017)'s LSTM in a replication of Malouf (2017)'s experiments using their original data from Flexique (Bonami et al., 2013). Flexique is an open source database for studying French inflection that builds on Lexique version 3.70 (New et al., 2001, 2004), an open source lexical database of French annotated with phonological, morphological, and frequency information. Lexique data is drawn from texts published after 1950 and subtitle files of French films available on the web and thus would not be expected to contain Old French.

In Malouf (2017)'s experiments (as well as Silfverberg and Hulden (2018)'s replications thereof) accuracy isn't only near-perfect for French (99.92%), but also highest for French out of 7 languages for both the LSTM system and the non-neural baseline from the 2017 SIGMORPHON shared task (99.06%) (Malouf, 2017, Table 2), described in §3.2.2. High accuracy on French is not due to the particular LSTM system, since the non-neural baseline did as well, and since the LSTM system did not perform well as the baseline in Silfverberg and Hulden (2018).

2.3 Hypotheses

In sum, in all but one of the inflection tasks reviewed in this section where UniMorph was the source of the French data, French accuracy was anomalously low relative to other languages. The one exception is Silfverberg and Hulden (2018)'s task, where the French UniMorph data was filtered to include only high frequency forms. When the source of French data was Flexique rather than UniMorph, accuracy was near perfect for both neural and non-neural models. In addition, unlike in the 2023 shared task, neural models outperformed the non-neural baseline on French in 2017 and 2018.

We hypothesized that: (i) Old French forms were prevalent in the UniMorph task splits when French yielded poor performance, i.e., in the 2017, 2018, and 2023 SIGMORPHON-UniMorph shared tasks, as well as Silfverberg and Hulden (2018)'s experiment that randomly sampled 1,000 inflection tables from French UniMorph, and (ii) Old French forms were not as prevalent or even absent in the task splits where French yielded better or near-perfect performance, i.e., in Silfverberg and Hulden (2018)'s experiment that filtered UniMorph inflection tables for high frequency forms, and in Malouf (2017)'s tasks splits from Lexique.

We also hypothesized that (iii) the prevalence of Old French forms in task splits was what was causing the anomalously poor performance on French. To support this hypothesis, we conducted three experiments replicating the 2023 SIGMORPHON-UniMorph task on French with the non-neural and neural baselines, first using the original task splits, and then re-sampling the task splits to exclude Old French lemmas in two different ways. Our prediction was that removing the Old French verbs from the task would lead to improvement in accuracy across both baseline models due to the elimination of errors related to Old French. We did not have a hypothesis about why the neural models failed to outperform the non-neural baseline on French in the 2023 shared task, but hoped that conducting an error analysis would reveal some insights.

3 Materials and methods

All source data, scripts for processing data, output files, and the error analysis spreadsheet can be found at `https://github.com/Prophecy0ak/TIGRE-2023sigmorphon`, which includes a README explaining how to run the scripts. The script `reproduce.sh` can be run to repeat the steps used to produce the output data from the SIGMORPHON-UniMorph 2023 shared task replication experiments used for error analysis.

3.1 Identifying Old French lemmas

3.1.1 Data

We checked the prevalence of of Old French lemmas in the French UniMorph 4.0 files `fra`, `fra.args` and `fra.segmentations`.[3] We also checked for Old French in

[3] `https://github.com/unimorph/fra`, accessed March 4, 2024

the following train/dev/test splits from past SIGMORPHON-UniMorph shared tasks: **2017/2018**: `french-train-high`, `french-dev`, `french-covered-test`[4]; **2019**: `french-train-high` from the french-occitan training data[5]; (iv) **2023**: `fra.trn`, `fra.dev`, `fra.test`[6].

The French UniMorph data from 2017 (UniMorph 1.0, (Kirov et al., 2016)) and 2018 (UniMorph 2.0 (Kirov et al., 2018)) included 7,535 lemmas and 367,732 forms. 12,000 triples were sampled without replacement for the splits. We only included the high-resource training data set from 2017 and 2018 (5,592 lemmas / 10,000 forms), since the medium and low resource training data were proper subsets of the high resource training data (Cotterell et al., 2017, Table 3; Cotterell et al., 2018, Table 2) .

In addition, we also checked for Old French lemmas in Malouf (2017)'s French data extracted from Flexique (`french.dat`[7]) and Silfverberg and Hulden (2018)'s random sample of 1000 lemmas from French UniMorph 2.0 (`fr.um.V.txt`) and sample of 1,131 lemmas filtered to contain high frequency forms (`fr.um.V.top.txt`).[8] Goldman and Tsarfaty (2021) reported using training/testing splits from Silfverberg and Hulden (2018).

3.1.2 Detecting Old French Lemmas

We determined which lemmas were Old French by writing a script `lang-stats.py` that checked the entry of each lemma in the English edition of Wiktionary for Old French.

Because this script relies entirely on Wiktionary entries, the definition of Old French may not be entirely accurate in all cases. While most pages cite *Dictionnaire de l'ancienne langue française et de tous ses dialectes du IXe au XVe siècle* (Godefroy, 1881) as the source for Old French definitions, not all pages had corresponding entries in said dictionary, so the reliability of Wiktionary for performing this task is questionable; however, given that we had no expertise in Old French, we chose to use the available Wiktionary entries in order to automate the process of checking the UniMorph data.

3.2 2023 SIGMORPHON-UniMorph replication experiments

3.2.1 Generating splits without Old French lemmas

To generate two new sets of splits without Old French lemmas from the original splits for the 2023 shared task, we filtered `fra.trn`, `fra.dev`, and `fra.tst` from the original splits using `fra.segmentations` in the current French UniMorph 4.0 repository. The `fra.segmentations` file contains morpheme segmentations developed for UniMorph 4.0 (Batsuren et al., 2022). We first confirmed that `fra.segmentations` contained no Old French lemmas using the procedure specified in §3.1.2. We then wrote a script `formatSegmentations.py` that converted `fra.segmentations` from the old feature schema from UniMorph 3.0 to the new hierarchical feature schema of UniMorph 4.0 used in the 2023 task splits. This new file `fra.total` was then sampled to create two sets of splits.

The **form-sampled** (seg-minimal) splits included only the *forms* that were contained in both the original splits and `fra.segmentations`; these splits were thus smaller than the original ones. The **lemma-sampled** (seg) set included only *lemmas* that were contained in both the original 2023 splits and `fra.segmentations` but all forms for those lemmas contained in `fra.segmentations`. Since `fra.segmentations` included many more forms than `fra`, the lemma-sampled splits were larger than the original splits. We based our splits on the original splits to preserve the original demographics, but wanted to account for both larger training and lower training amounts without adding in too much of our own biases.

3.2.2 Algorithms

Since one anomalous aspect of performance on French in the 2023 shared task was that the non-neural baseline outperformed neural models, we included both the non-neural baseline[9] and the neural baseline (Wu et al., 2021)[10] in our experiments. The non-neural baseline has been used

[4] https://github.com/sigmorphon/conll2017/tree/master/all/task1, https://github.com/sigmorphon/conll2018/tree/master/task1/surprise

[5] https://github.com/sigmorphon/2019/blob/master/task1/french--occitan/french-train-high

[6] https://github.com/sigmorphon/2023InflectionST/tree/main/part1/data

[7] https://github.com/rmalouf/abstractive/blob/master/data/french.dat.gz

[8] https://github.com/mpsilfve/pcfp-data/tree/master/data

[9] Accessed from https://github.com/sigmorphon/2023InflectionST/blob/main/part1/baselines/nonneural.py

[10] Accessed from https://github.com/omagolda/neural-transducer/tree/master/example/sigmorphon2023-shared-tasks

for SIGMORPHON-UniMorph shared tasks since Cotterell et al. (2017), and the neural baseline, a character-level transformer, since Pimentel et al. (2021). The non-neural records prefixing and suffixing rules, and then uses a matching heuristic to decide which rules to apply given the set of features.

We did not test systems other than the baselines, since the focus of this paper is issues with the gold data independent of algorithm choice. Also, code was not yet available for Canby and Hockenmaier (2023)'s top-performing neural systems; the other neural system was outperformed by the neural baseline anyway, and the submitted finite state transducer systems performed comparably to the non-neural baseline.

3.2.3 Error taxonomy

The error taxonomy we used is an extension of Gorman et al. (2019)'s annotation scheme for the 2017 SIGMORPHON-UniMorph shared task. Gorman et al. (2019, p. 142) split errors into four major categories, three of which are cited below and used in an identical fashion. We omitted spelling errors due to a lack of errors that differed only in spelling.

Since we were trying to account for the influence of Old French verbs in a given error, we added a superordinate category **Old French errors**.

Old French errors This category includes all errors that can be attributed to the presence of outdated verbs in training, development, and test data. We specified two sub-categories: (i) **Old French Lemma errors**, for Old French verb lemmas that are not used in modern French, i.e., extraction errors, and (ii) **Old French affix overapplication errors**, which involve applying Old French inflecting patterns learned from muddied data to modern verbs. These were not considered allomorphy errors because the Old French affixes did not constitute "existing allomorphic patterns in the target language" (Gorman et al., 2019), i.e. French.

Free Variation This category was the same as Gorman et al. (2019)'s free variation category and included verbs which have "free variation" in French, but where only one form was available due to the UniMorph scraping procedure. In these cases, the error was a grammatical form but not included as a correct form in the gold data.

Allomorphy Errors These were divided into two subcategories used in Gorman et al. (2019) but not reported in the paper[11]: (i) **Affix overregularization errors**: errors where the target irregular affix was replaced with one that is regular. (ii) **Affix overirregularization errors**: errors where regular affixes were replaced by irregular affixes.

Silly Errors This category was the same as Gorman et al. (2019)'s silly error category and encompassed cases where the model's prediction was extremely dissimilar to the gold data. This dissimilar form was not present elsewhere in the given inflectional category for the language. Silly errors included completely strange and random inflectional forms that differed greatly from the lemma and were primarily seen in inflection errors made by the neural model, see §4.2.1.

3.2.4 Annotation procedures

The annotation conducted for Gorman et al. (2019)'s experiment was annotated by native speakers and some by second-language speakers with expertise in computational linguistics. Our annotators fall into this second category, thus, annotation of the error data was carried out both as researchers with backgrounds in linguistics and as advanced French speakers.

Our error categorization used an order of priority similar to Gorman et al. (2019)'s, though starting with **Old French errors** and proceeding thereafter through **Free Variation**, **Allomorphy**, and **Silly**. However, the first step of this priority order was only applied very conservatively. The **Old French Lemma** error category was only applicable to those lemmas which, according to Wiktionary, were not modern French verbs. The **Old French affix overapplication** error category was only applied when we found the model had used a suffixing rule which exists nowhere in modern French but had been scraped from an Old French Wiktionary entry.

4 Results

4.1 Prevalence of Old French in past inflection tasks and UniMorph files

4.1.1 UniMorph 4.0 files

The `fra.args` file[12]—which seems to the source file for 2023 SIGMORPHON-UniMorph shared task splits—contained 20.8% (1564/7535) lemmas from Old French, and we confirmed that there were no old lemmas in `fra.segmentations`.

[11] Thanks to Kyle Gorman sharing full annotation scheme.

[12] and also the `fra` file, which is different from `fra.args` only in using the UniMorph 3.0 feature scheme

Year	Frequency of Old lemmas			Freq. of inflected forms from Old lemmas		
	Train	Dev	Test	Train (/10,000)	Dev (/1,000)	Test (/1,000)
2017	1,146 (20.5%)	182 (19.4%)	193 (20.5%)	2,045 (20.4%)	194 (19.4%)	206 (20.6%)
2018	1,165 (20.8%)	214 (22.6%)	203 (21.6%)	2,108 (21.1%)	221 (22.1%)	215 (21.5%)
2019	1,139 (20.5%)	N/A	N/A	2,052 (20.5%)	N/A	N/A
2023	86 (21.5%)	5 (10%)	9 (18%)	2,142 (21.4%)	100 (10%)	180 (18%)

Table 1: Raw and relative frequencies of Old French lemmas and forms inflected from Old French lemmas in SIGMORPHON-UniMorph shared task splits. Only high-resource training sets were included, see §3.1.1 for details.

4.1.2 SIGMORPHON-UniMorph shared task splits

We determined that Old French lemmas typically occurred in approximately 20% of each of the train/test/dev splits in past SIGMORPHON-UniMorph shared tasks involving French, as summarized in Table 1.

4.1.3 Minimal supervision inflection tasks

We found that Malouf (2017)'s French data extracted from Flexique contained a very small number (7 out of 5220) of Old French lemmas (see §5). This was unexpected since Flexique is based on post-1950s texts and subtitles from French movies. Additionally, Silfverberg and Hulden (2018)'s random sample of 1000 lemmas from French UniMorph 2.0 (`fr.um.V.txt`) contained 216 Old French lemmas (21.6%) while the high-frequency filtered sample (`fr.um.V.top.txt`) included only 114 historical lemmas (10.1%).

4.2 2023 shared task replication experiments

The number of distinct lemmas and forms for each split for all three experiments is given in Table 2. Filtering the original splits via `fra.segmentation` resulted in the loss of 22%, 10%, and 18% of lemmas in each split respectively. The form-based split ended up having 3 fewer lemmas because the matching performed in our script does not account for the inconsistencies in `fra.segmentations` and `fra` in representing reflexive forms.

Removing Old French lemmas improved the accuracy for the non-neural baseline ("RU") by 10.32-11.72% and for the neural ("NN") algorithm by 12.32-13.34% (Table 3). Whether the original splits were re-sampled by-lemma ("Seg") or by-form ("Seg-Minimal") made only about a 1% difference. Even with the re-sampled splits excluding Old French lemmas, the non-neural baseline still outperformed the neural baseline by about 10%.

Split	Original	Seg	Seg-Minimal
Train	400:10,000	312:15,890	309:6,407
Dev	50:1,000	45:2,265	45:754
Test	50:1,000	41:2,101	41:701

Table 2: Number of distinct lemmas:forms in each split for each experiment.

System	Original	Seg	Seg-Minimal
RU	83.15%	94.87%	93.47%
NN	71.40%	83.72%	84.74%

Table 3: By-form accuracy for the non-neural (RU) and neural (NN) models, aggregated across test and dev sets.

4.2.1 Error analysis

The distribution of error types (defined in §3.2.3) for each of the three experiments (original, Seg by lemma, Seg-Min by form) and algorithms is shown in Figure 1, combining errors across dev and test sets. Table A1 shows the raw counts for error types in dev and test sets separately. The abbreviations in the figure and table correspond to the error taxonomy categories as ordered in §3.2.3.

Old Lemma and Old French affix overapplication errors ("**Old Rule**") occurred only for the original splits, comprising 56.6% and 37.8% of errors for the non-neural and neural models, respectively. Example overapplication errors are in Table A2. For both models, the other major error type was affix overregularization ("**Over Reg**") allomorphy errors, comprising 33-37% of the errors for the original splits. Overregularization was the most frequent error in the Seg and Seg-Min resampling experiments—about 90% of errors for the nonneural model and 56-64% of errors for the neural model. The neural model differed from the nonneural model in having many silly errors—20% of the errors for the original splits and 32-40% for the Seg-resampled splits.

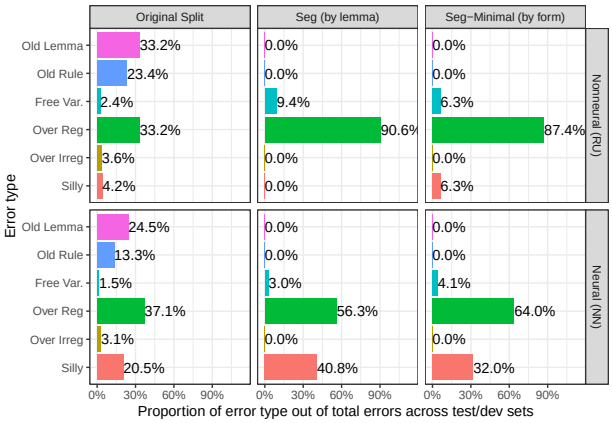

Figure 1: Proportion of error type out of all errors across test/dev sets for neural (NN) and non-neural (RU) baselines for each experiment.

5 Discussion

We determined that Old French comprised around 20% of the lemmas in the SIGMORPHON-UniMorph 2017, 2018, and 2023 shared task splits, as well as Silfverberg and Hulden (2018)'s random sample of 1,000 inflection tables from French UniMorph. These were all cases when French yielded anomalously poor performance relative to other languages. However, Silfverberg and Hulden (2018)'s sample of French UniMorph filtered for high frequency forms, which yielded better performance, only had 10% Old French lemmas, and Malouf (2017)'s French data from Lexique that yielded near-perfect accuracy had only 0.13% Old French lemmas.

In short, performance on French inflection tasks was inversely proportional to the proportion of Old French lemmas present in the task data. Furthermore, errors related to the presence of Old French in the data were prevalent in our replication of the 2023 shared task with the original task splits for both non-neural and neural models. Removing Old French from the splits eliminated these errors. Interestingly, the non-neural baseline still outperformed the neural baseline even when Old French lemmas were removed. Thus, the presence of Old French in the original 2023 task splits doesn't seem to be the cause of the the non-neural baseline outperforming the neural model.

These improvements suggest that correctly separating modern, Old, and Middle French into separate datasets is important for computational morphology tasks. UniMorph itself has separate repositories for Old (`unimorph/fro`) and Middle (`unimorph/frm`) French, so the erroneous inclusion of lemmas from both Old and Middle French creates confusing inconsistencies for potential future projects which may want to work on all three languages.

This bug was noted and addressed in the UniMorph 3.0 revision (McCarthy et al., 2020): "Finally, a bug in the previous extraction process caused languages' data to be read into other languages' files whose names are their suffixes. For instance, 'Greek' contained data from 'Ancient Greek', and 'French' contained data from 'Middle French'. Filtering and rerunning our extraction process eliminated these erroneously grouped paradigms" (McCarthy et al., 2020, p. 3924). However, the issue persists in the `fra` and `fra.args` data files in UniMorph 3.0 and UniMorph 4.0.

It is worth noting that the task of distinguishing Old and modern French involves a degree of nuance. The sampled Flexique data contained 7 lemmas which were classified as Old French by our script, but according to the French edition of Wiktionary, these words have been repurposed as either idiomatic expressions or legal terms in modern French, now using modern inflection patterns. While future projects should be sure that their data makes this distinction, simple scraping of the English edition of Wiktionary may present issues for obtaining truly representative lexical data.

5.1 Data Inconsistencies

Elsner et al. (2019, p. 78-79) notes that none of the SIGMORPHON datasets provide an adequate lexical set to account for the Zipfian distribution of words in natural language. For example, "spotty coverage of high frequency words for German appears to be typical of the UniMorph datasets." Similarly, we found that our splits lacked highly frequent verbs such as *être* ('to be'), *faire* ('to do'), and *pouvoir* ('to be able to'), which were included in the more exhaustive `fra.segmentations`. Despite the limited size of the training data, we nevertheless noticed some further data inconsistencies that would have caused more issues if they had been included in the dataset to the extent that they are represented in the language. This includes (i) inconsistency in the documentation of French reflexives in Wiktionary, and (ii) the presence of multiple possible grammatical inflections for verbs such as *-eler* and *-eter* verbs.

5.1.1 Reflexive inconsistencies

Reflexive verbs in French include a reflexive pronoun *se* (onesself) that is the object of the verb, e.g. *il se regarde* ('he looks at himself'). Despite there being no reflexive verbs in the test or development splits, reflexives verb forms are quite common in the French language. The three lemmas that did appear with reflexive pronouns in the training split were inconsistently recorded (two had reflexives pronouns in the inflected forms but not in the lemma, while the third had a reflexive pronoun in the lemma as well). These inclusions were enough to cause the neural model to erroneously identify *génuflexionner* ('to bend the knee') as a reflexive verb, though this verb takes no object.

Had more reflexives been included in the train and test splits, the effect of inconsistent data on the models' accuracies would have been much greater. These inconsistencies include duplicate pages, transitive verb pages with "reflexive" usage shown in the definitions but not in the conjugations, and those listed as transitive but conjugated using reflexive pronouns. Many of the most common reflexive verbs are entirely missing or have been deleted due to differing opinions on the necessity of the reflexive form having separate documentation. Had they been included in our splits, we predict that the inconsistencies would have posed issues for properly measuring each model's performance on French.

5.1.2 Multiple grammatical inflections

There exists a prescriptivist body in the French government, l'Académie Française, which is tasked with publishing the French dictionary as well as setting official orthography changes in the language over time. This has resulted in a degree of free variation in the inflection of French verbs. In accordance with the Académie's prescriptions, Wiktionary has a number of French verb charts that have multiple options mapped to a single morphosyntactic tag, where UniMorph only scrapes one option per lemma/feature pair. The most common of these are *-eler* and *-eter* verbs, which can now be conjugated by either doubling the consonant or adding an è before said consonant, except for those derived from *appeler* ('to call') or *jeter* ('to throw').

Since UniMorph only scrapes one option, when models predict one of the other permitted conjugations, they are marked incorrect. There was only one of the aforementioned *-eler* and *-eter* verbs in our data, *craqueler* ('to crack') in the dev split. The errors that resulted from this free variation were noted in the annotation scheme, but such errors would be much greater in number if the data had been more exhaustive. By performing a more inclusive scrape of Wiktionary that grabs all of the grammatical inflected versions of a lemma with a given morphosyntactic tag,[13] we predict there would be an increase in accuracy since this would mark inflections correct that would previously have been erroneously marked as a mistake in the predicted form.

5.2 Proposed fixes for French

We propose that future shared inflection tasks use `fra.segmentations` rather than `fra/fra.args`, which would eliminate all errors that fell under the **Old French error** category in our taxonomy. The improvements to accuracy as a result of this change are reflected in our results. Using `fra.segmentations` instead would also allow more comprehensive inclusion of common French verbs, which would generate results that are more reflective of how these models handle the French lexicon.

Additionally, we advise caution in scraping French reflexive verbs from the English edition of Wiktionary, as well as verbs with free variation, as described in §5.1. Wiktionary is subject to inconsistencies as well as disagreement between Wiktionary entry authors despite its richness in linguistic data.

Finally, as French is a very well-documented language, there are several other resources for linguistic data which may be more consistent and reliable than the English edition of Wiktionary. These could help circumvent data consistency issues in future computational linguistics tasks. For example, the Morphalou3 lexicon takes into account purely orthographic variations on individual words, including those allowed by the additional rules prescribed by the French Academy in 1990, and is a consolidation of Morphalou with 4 other French lexicons (DELA, Dicollecte, LGLex/LGLexLefff, and Lefff) (ATILF, 2023). The GLÀFF lexicon (Hathout et al., 2014) is specifically based on the French edition of Wiktionary and thus does not have the same consistency issues as the English edition. It also includes the overall frequency of each lexeme (per million

[13] Malouf et al. (2020, §3.4) has an alternative suggestion: to remove paradigms with multiple grammatical inflections from the data.

words across various large French corpora), which would be helpful in selecting for more common words when designing train/test/dev sets.

5.3 Beyond French

It was only because of our in-depth attention to French results and our particular linguistic knowledge of French that we were able to spot the erroneous inclusion of Old French in the UniMorph data and then perform the qualitative error analyses in this paper. There are, no doubt, other UniMorph languages which could benefit from similar language-specific studies. Yet, while the cross-linguistic coverage of UniMorph and SIGMORPHON-UniMorph shared tasks has rapidly expanded across the past decade, detailed, language-specific analyses of UniMorph data and/or SIGMORPHON-UniMorph results remain few in number.

Studies that have examined particular languages in detail have found issues with Wiktionary data and/or extraction errors—for instance, in Romanian, Hungarian, and Latin (3 of the 12 languages examined in the error analyses of Gorman et al. (2019)), as well as Navajo (Malouf et al., 2020). In an examination of UniMorph data, Malouf et al. (2020) raises some of the same issues that we discussed in §5: limited size of data sets and the availability of multiple grammatical inflectional forms for a single paradigm cell. Malouf et al. (2020) points out that there are several inconsistencies in choices made by Wiktionary editors for Navajo entries which negatively affect the overall performance of morphological inflection models when using Navajo data from UniMorph. For instance, Wiktionary provides separate entries for bare nouns and their possessed forms for some but not all Navajo lemmas. While the possessed forms should certainly be included, the decision to keep the entries separate for certain nouns is confusing and causes some inflected forms to be treated as lemmas in their own right.

6 Conclusion

When shared tasks include dozens and dozens of languages, it is hard to interpret results when each individual language could be affected by data issues like those we have discussed in this paper. Such problems underscore the need for shared tasks to include qualitative, language-by-language analysis of data and results in addition to reporting accuracy. It is admittedly a tall order to do analyses like the one in this paper for each of the over two dozen languages from the SIGMORPHON-UniMorph 2023 shared task, but perhaps shared tasks could explicitly focus on probing and improving data quality and otherwise emphasize language-by-language error analysis as an essential step of analyzing results. This kind of work would naturally encourage collaboration between language experts/linguists and modelers, as suggested in Malouf et al. (2020)'s statement of best practices for computational modeling of cross-linguistic morphology. By closely examining the distribution of errors produced, future projects can concentrate on eliminating prevalent error categories that have previously hindered model performance, enabling focused improvements in shared tasks.

Acknowledgements

We thank Kyle Gorman for discussion of Gorman et al. (2019), UniMorph mailing list (https://groups.google.com/g/unimorph) members for responding to questions about French UniMorph, and members of the UMass Amherst LINGUIST 409 Introduction to Computational Linguistics Fall 2023 course for helpful feedback.

References

Rami Al-Rfou', Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed Word Representations for Multilingual NLP. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria. Association for Computational Linguistics.

ATILF. 2023. Morphalou. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

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 Conference*, pages 840–855, Marseille, France. European Language Resources Association.

Olivier Bonami, G. Caron, and C. Plancq. 2013. Flexique : An inflectional lexicon for spoken French.

Marc Canby and Julia Hockenmaier. 2023. A Framework for Bidirectional Decoding: Case Study in Morphological Inflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4485–4507, Singapore. Association for Computational Linguistics.

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.

Micha Elsner, Andrea D. Sims, Alexander Erdmann, Antonio Hernandez, Evan Jaffe, Lifeng Jin, Martha Booker Johnson, Shuan Karim, David L. King, Luana Lamberti Nunes, Byung-Doh Oh, Nathan Rasmussen, Cory Shain, Stephanie Antetomaso, Kendra V. Dickinson, Noah Diewald, Michelle McKenzie, and Symon Stevens-Guille. 2019. Modeling morphological learning, typology, and change: What can the neural sequence-to-sequence framework contribute? *Journal of Language Modelling*, 7(1):53–98.

Frédéric Godefroy. 1881. *Dictionnaire de l'ancienne Langue Française et de Tous Ses Dialectes Du IXe Au XVe Siècle*. F. Vieweg, Paris.

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. 2021. Minimal Supervision for Morphological Inflection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2078–2088, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kyle Gorman, Arya D. McCarthy, Ryan Cotterell, Ekaterina Vylomova, Miikka Silfverberg, and Magdalena Markowska. 2019. Weird Inflects but OK: Making

Sense of Morphological Generation Errors. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 140–151, Hong Kong, China. Association for Computational Linguistics.

Nabil Hathout, Franck Sajous, and Basilio Calderone. 2014. GLÀFF, a Large Versatile French Lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1007–1012, Reykjavik, Iceland. European Language Resources Association (ELRA).

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina J. Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. UniMorph 2.0: Universal Morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Christo Kirov, John Sylak-Glassman, Roger Que, and David Yarowsky. 2016. Very-large Scale Parsing and Normalization of Wiktionary Morphological Paradigms. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3121–3126, Portorož, Slovenia. European Language Resources Association (ELRA).

Robert Malouf. 2017. Abstractive morphological learning with a recurrent neural network. *Morphology*, 27(4):431–458.

Robert Malouf, Farrell Ackerman, and Arturs Semenuks. 2020. Lexical databases for computational analyses: A linguistic perspective. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 446–456, New York, New York. 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.

Boris New, C. Pallier, Ludovic Ferrand, and Rafael Matos. 2001. Une base de données lexicales du français contemporain sur internet : LEXIQUE™//A lexical database for contemporary french : LEXIQUE™. *L'année psychologique*, 101(3):447–462.

Boris New, Christophe Pallier, Marc Brysbaert, and Ludovic Ferrand. 2004. Lexique 2 : A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, 36(3):516–524.

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.

Miikka Silfverberg and Mans Hulden. 2018. An Encoder-Decoder Approach to the Paradigm Cell Filling Problem. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2883–2889, Brussels, Belgium. Association for Computational Linguistics.

Shijie Wu, Ryan Cotterell, and Mans Hulden. 2021. Applying the Transformer to Character-level Transduction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1901–1907, Online. Association for Computational Linguistics.

Appendix

	Old Lemma	Old Rule	Free Var.	Over Reg	Over Irreg	Silly	Total
Orig/RU/dev	33	31	8	34	8	0	114
Orig/NN/dev	30	30	7	64	8	50	189
Orig/RU/test	79	48	0	78	4	14	223
Orig/NN/test	82	31	0	106	6	44	269
Seg/RU/dev	0	0	21	41	0	0	62
Seg/NN/dev	0	0	21	123	0	125	269
Seg/RU/test	0	0	0	162	0	0	162
Seg/NN/test	0	0	0	277	0	165	442
Min/RU/dev	0	0	6	18	0	1	25
Min/NN/dev	0	0	9	46	0	52	107
Min/RU/test	0	0	0	65	0	5	70
Min/NN/test	0	0	0	96	0	19	115

Table A1: Frequency of errors types for dev and test splits for experiment (original, Seg, Seg-Minimal) and algorithm (RU vs. NN). The errors are listed left to right in the order of taxonomy priority.

lemma	features	gold	model prediction
absoudre	COND.3SG	absoudrait	absoudr**oit**
	IND.PST.PFV.2PL	absolûtes	absou**istes**
désarmer	COND.2SG	désarmerais	désarmer**ois**
	IND.PST.IPFV.3SG	désarmait	désarm**oit**
	IND.PST.PFV.1PL	désarmâmes	désarm**asmes**
délayer	IND.PST.IPFV.3PL	délayaient	délay**oient**
	IND.PST.IPFV.1SG	délayais	délay**ois**
alanguir	IND.PST.PFV.1SG	alanguis	alangu**a**
abonder	IND.PST.PFV.2PL	abondâtes	abond**astes**
mendier	SUBJ.PST.3SG	mendiât	mend**ast**
tuner	SUBJ.PST.2PL	tuneriez	tuniss**oiz**
objectiver	SUBJ.PRES.2PL	objectiviez	objectiv**ez**

Table A2: Examples of modern French verbs erroneously inflected with Old French suffixes. Triples mentioned in §1 in first three columns, fourth column is an error that falls into the **Old French affix overapplication** (**Old Rule**) category. Refer to yellow-highlighted data in `ErrorAnnotations.xlsx` in the GitHub repository.

The effect of model capacity and script diversity on subword tokenization for Soranî Kurdish

Ali Salehi **Cassandra L. Jacobs**
Department of Linguistics
University at Buffalo
asalehi;cxjacobs@buffalo.edu

Abstract

Tokenization and morphological segmentation continue to pose challenges for text processing and studies of human language. Here, we focus on written Soranî Kurdish, which uses a modified script based on Persian and Arabic, and its transliterations into the Kurdish Latin script. Importantly, Perso-Arabic and Latin-based writing systems demonstrate different statistical and structural properties, which may have significant effects on subword vocabulary learning. This has major consequences for frequency- or probability-based models of morphological induction. We explore the possibility that jointly training subword vocabularies using a source script along with its transliteration would improve morphological segmentation, subword tokenization, and whether gains are observed for one system over others. We find that joint training has a similar effect to increasing vocabulary size, while keeping subwords shorter in length, which produces higher-quality subwords that map onto morphemes.

1 Introduction

Different scripts for the same language may convey different linguistic and structural properties, such as phonological transparency, word boundaries (e.g., whitespace), morpheme boundaries, serial position within a word, or present different orthotactic and spelling constraints. In this work, we examine the relationship between script variation, morphological acquisition, and subword vocabulary construction. Obtaining high-quality morphological annotations is critical for linguistic analysis, so unsupervised methods for learning morpheme-like representations are often an acceptable compromise. Here, we explore the usefulness of subword vocabulary training for morphological segmentation of written Soranî Kurdish, a central dialect of Kurdish spoken mainly in Iran and Iraq. Soranî is morphologically complex but relatively underresourced, with few large annotated corpora (Veisi et al., 2019; Malmasi, 2016; Goldhahn et al., 2012; Ahmadi, 2020a; Mahmudi and Veisi, 2021), and none with adequate morphological glosses or segmentations for downstream language model development (Alkaoud and Syed, 2020; Banerjee and Bhattacharyya, 2018) or linguistic analysis.

Soranî has some unlabeled raw text corpora, which opens the possibility to leverage the statistical properties of the text for unsupervised subword-based vocabulary induction. The existence of multiple writing systems for Kurdish languages additionally presents a challenge for NLP systems, and jointly training subword tokenization models may be advantageous for Central Kurdish NLP in general. We thus ask whether training a subword vocabulary on multiple scripts can induce adequate morphological segmentations and compare such systems against models trained solely on single scripts of equivalent or larger sizes.

2 Soranî morphology and script variation

Our manipulation leverages script variability in the written Kurdish dialects. Kurdish dialects have been written with diverse writing systems including Arabic, Latin, Yekgirtú (unified), Cyrillic and Armenian scripts. There is no unified orthography for Kurdish despite previous efforts (Ahmadi et al., 2020). This variation presents an intriguing opportunity to explore the impact of input diversity on the learning of subword vocabulary. Furthermore, demonstrating the potential usefulness of joint training on multiple scripts could produce higher-quality multi-dialect Transformer language models (Kanjirangat et al., 2023).

The Soranî writing system used for Central and Southern Kurdish is written in a modified Perso-Arabic script and has an alphabetic structure with a high degree of phonological transparency, relative to Arabic and Persian scripts (Chyet and Schwartz, 2003; Ahmadi, 2020b). The Latin-based Hawar al-

Proceedings of the The 21st SIGMORPHON workshop on Computational Morphology, Phonology, and Phonetics, pages 51–56

phabet, used by Northern Kurdish dialects, shares this transparency, making it feasible to transliterate Soranî script into a Latin-based one (Mahmudi and Veisi, 2021). The Latin script has two allographs per segment (e.g., H/h), which mostly encode sentence position, but the Perso-Arabic script has three for word-initial, -medial, and -final positions (e.g., the phoneme /h/ is represented by 'هـ' word initially, by 'ـهـ' word medially and by 'ه' or 'ـه' word finally.

3 Subword tokenization

We explore multiple tokenization models for Soranî Kurdish, with a primary focus on Byte Pair Encoding (BPE; Sennrich et al., 2016a) and Unigram tokenization (Kudo, 2018). We assessed these models' performance with respect to morphological and phonological structure and critically assess claims of better morphological induction by Unigram relative to BPE (Bostrom and Durrett, 2020). Both BPE and Unigram LM have probabilistic components based on frequency and vocabulary likelihood, respectively. These models were selected for their ability to handle diverse linguistic data and to learn meaningful linguistic units from large datasets without extensive annotated resources. The focus on these tokenization models that are used in modern neural methods is to align any prospective tokenization system with current trends, enabling scalability and robustness across different datasets and providing an exploration of their adaptability to the morphological richness of Kurdish in both Latin and Arabic scripts.

Byte-Pair Encoding. BPE is a simple and common subword tokenization algorithm (Gage, 1994; Sennrich et al., 2016a) that grows a vocabulary from individual characters into more complex subwords by merging the most frequent co-occurring character sequence, up to a specified number of merges. Given the frequency-based merging process of BPE, it is plausible that manipulating the relative frequency of subwords by training multiple scripts will influence the resulting vocabularies. We further hypothesize that the different character frequency distributions of Latin and Perso-Arabic scripts may help BPE to learn subwords that better align with morphological boundaries and better capture the tendency of non-stem morphemes in Kurdish to be short.

Unigram tokenization. This subword method iteratively splits words into subwords by optimizing the likelihood of the training data, which providing a probabilistic approach to subword segmentation that may capture more nuanced linguistic patterns compared to BPE's frequency-based merging strategy (Kudo, 2018). Unigram tokenization has been argued to produce better morphological segmentations than algorithms like BPE or WordPiece (Bostrom and Durrett, 2020). We expect Unigram tokenization to potentially provide more comprehensive coverage of Soranî morphology compared to BPE, due to the likelihood objective of Unigram.

4 Experiments

We used the huggingface tokenizer package for BPE (Sennrich et al., 2016b) and Unigram tokenization (Kudo and Richardson, 2018). For our experiments, we used the normalized version of the Asosoft corpus (Veisi et al., 2019) consisting of 188 million word tokens and 4.66 million word types. The corpus was chosen for its comprehensive coverage of the Soranî dialect. The corpus includes 58,000 documents from textbooks and magazines and 400,000 documents from web crawls. We removed newline characters, repeated characters (Rajadesingan et al., 2015), and redundant whitespace before subword training. We tested separate models for Latin and Arabic scripts, each with a 5k vocabulary size. Additionally, we constructed a joint Arabic-Latin script corpus for data augmentation and further constrained model size based on script-specific vocabulary sizes derived from this joint corpus.

Vocabulary size. We explored various vocabulary sizes within a range of 1,000 to 10,000 subwords to identify the optimal balance between granularity and generalization. For BPE, we found that a larger vocabulary size of 5,000 subwords provided the best results, and so we use this size in all our experiments. This size qualitatively offered a good trade-off between identifying roots and affixes versus learning morphologically complex words, capturing the morphological complexity of Soranî in both Latin and Arabic scripts. Across all of our measures, 5,000 subwords each for the Latin and Arabic scripts led to the highest performance.

Transliteration. The Latin-based script exhibits a one-to-one correspondence between phonemes and alphabet letters (Esmaili et al., 2013) that can be deterministically transliterated from the Arabic script using Asosoft (Mahmudi and Veisi, 2021). In addition to changing character frequencies caused

Model	Vocab Size	Script	Avg. Len.	Tokenization Agreement (%)	Syllabification (%)
BPE - Small	2514	Latin	2.94	75.29	
	2446	Arabic	2.87		4.87
BPE - Large	5000	Latin	3.46	**79.67**	
	5000	Arabic	3.38		12.64
BPE - Joint	2514	Latin	3.04	77.08	
	2446	Arabic	2.98		**26.70**
Unigram - Large	5000	Latin	3.33	74.28	
	5000	Arabic	3.24		11.73
Unigram - Joint	3892	Latin	3.09	76.72	
	3647	Arabic	3.01		25.77

Table 1: Comparison of tokenization models for Soranî Kurdish in Latin and Arabic scripts.

by multiple allographs, transliteration into the Latin script introduces the letter "i" for the schwa, which is not encoded in the Perso-Arabic script. The relative transparency of the Latin script may produce more accurate segmentations than the Arabic script.

Data "augmentation." We define a joint tokenizer as a tokenization model trained on text data from multiple scripts simultaneously. For Kurdish languages, which can be written in both Latin and Arabic scripts, a joint tokenizer aims to create a unified subword vocabulary that can effectively tokenize text for multiple dialects. This approach combines text data from both scripts for a balanced training set, which the BPE and Unigram model then uses to develop a script-agnostic tokenization strategy based on subword frequency. This effectively doubles the training data set size and may alter the relative frequencies of subwords in the data. We measure the different tokenizers' precision against verified morphological segmentations of Soranî Kurdish, along with segmentation accuracy. We hypothesize changes in subword tokenization following from the fact that the two scripts have slightly different orthotactics (see Section 2). Transliteration is hypothesized to enhance subword vocabulary training by increasing the number of data points under consideration (Shazal et al., 2020; Biadgligne and Smaili, 2023).

5 Results

5.1 Subword vocabularies

We first characterize the subword vocabularies and their behavior for words in the training corpus. Our analysis includes the token match rate between Latin and Arabic scripts, average token length, syllable-token correspondence, and token-morpheme match rate to assess the effectiveness of subword tokenization models in capturing the linguistic structure of Soranî Kurdish (Table 1).

Token length. The average length of the tokens reveals the granularity of the subword segmentation, with shorter lengths indicating finer segmentation. Unigram model tends to produce longer subwords, indicating differences in the granularity of tokenization caused by the split procedure. The BPE models produce shorter subwords, which follows given the merge-based training procedure, and this is especially true for small, single-script models.

Tokenization consistency. The percentage of tokenizations that are at the same boundaries across both Arabic and Latin scripts (Token % in Table 1) highlights the models' ability to maintain consistency across different writing systems, which is crucial for script-agnostic NLP applications. It is meant to measure the consistency of tokenization by comparing boundary positions in both scripts, quantifying the percentage of boundaries that coincide. The larger independently-trained BPE models achieve the highest token match rate at 79.67%, suggesting that similar types of merges are occurring for both scripts.

Syllabification. Syllable-token correspondence measures the alignment of tokens with the syllabic structure of the language. The highest percentage of matching occurs with BPE trained jointly.

5.2 Morphological coverage

We assess the quality of the subword vocabularies by computing the overlap between the tokens generated by BPE and the morphemes of the words, as well as the proportion of token strings that cor-

Setting	Model	Mean tokens in morpheme set	Mean morphemes in vocabulary	Morpheme Coverage %	Segmentation Accuracy %
Latin Script					
Joint training	BPE	0.349	0.345	43	26
	Unigram	*0.428*	*0.423*	47	*34*
2514 subwords	BPE	0.336	0.333	41	25
5000 subwords	BPE	0.379	0.370	*50*	29
	Unigram	**0.440**	**0.435**	**51**	**36**
Arabic Script					
Joint training	BPE	0.368	0.361	44	28
	Unigram	*0.484*	*0.479*	49	*40*
2446 subwords	BPE	0.353	0.349	41	26
5000 subwords	BPE	0.402	0.390	*52*	32
	Unigram	**0.496**	**0.485**	**54**	**43**

Table 2: Performance metrics of tokenization models for Soranî Kurdish.

respond to morphemes and the proportion of morphemes that are present in the subword vocabulary.

To create the test set for evaluating the tokenization models, we selected words from the corpus that represent a variety of linguistic phenomena in Soranî. This included words with ezafe constructions, compounds, preverbal constructions, and words that incorporate prepositions. We also chose words that contain a half space or Zero Width Non-Joiner (ZWNJ) to assess the models' ability to handle this aspect of the script. To evaluate the models' performance in capturing the morphological structure of Soranî, 1500 words were manually tokenized to accurately segment the morphemes. Table 3 illustrates the efficacy of different tokenization models in segmenting Soranî words into their respective morphemes. We compare Unimorph and BPE with different vocabulary sizes, across a selection of words. The comparison focuses on how each model tokenizes the words and aligns these tokens with the linguistically motivated morpheme boundaries. For instance, the word "destîpêkird" is tokenized differently by Unimorph and BPE, reflecting each model's approach to parsing the underlying morphological structure of the language.

Combining two scripts has a small but positive effect on tokenization quality in terms of morphological accuracy for BPE, relative to the small single-script models. When BPE is trained to a larger subword vocabulary for either script, it performs slightly better in terms of morphological coverage compared to other models, including the joint BPE model. This highlights the potential trade-offs between vocabulary size and script coverage in subword tokenization for Soranî Kurdish. However, Unigram tokenization consistently outperforms on all measures of morphological structure, as seen in prior work (Bostrom and Durrett, 2020). We summarize these comparisons in Table 2.

6 Future work

A specific subset of words containing the Zero-width non-joiner (ZWNJ) was deliberately isolated to assess tokenization performance of the Unigram tokenizer with 5000 subwords, particularly within the Perso-Arabic script. The presence of ZWNJ, which can act as a morphological delimiter in word or appear after the letter 'ه' /a/ by pressing the E key on the keyboard, helps in achieving more accurate segmentation outcomes. For instance, in the word *tokmetir* 'تۆکمەتر' 'stronger', the word *tokme* 'تۆکمە' is separated from the comparative morpheme *tir* 'تر' by a ZWNJ which gets tokenized as a two actual morphemes by the Unigram tokenizer. This structural feature can provide clues to tokenization models, enabling more precise identification and segmentation of morphemes and a higher granularity in morpheme segmentation compared to the Latin script. Such findings underscore the importance of leveraging script-specific orthographic cues to improve tokenization models for

Tokenizer	vocab	word	Latin Tokens		Morphemes
Unimoph	5k	destîpêkird	[‘destîpêkird’]	[دەستپێکرد]	dest-î-pê-kird
Unimorph	5k	meseley	[’meseley’]	[مەسەلەی]	mesele-y
Unimorph	5k	pîlangêřîyekan	[‘pîlan’, ‘gêř’, ‘îyekan’]	[پیلان، گێڕ، سەکان]	pîlan-gêřî-yek-an
BPE	5k	destîpêkird	[‘destî’, ‘pêkird’]	[دەستی، پێکرد]	dest-î-pê-kird
BPE	5k	lebexda	[‘lebexda’]	[لەبەغدا]	le-bexda
BPE	2514/2446	damezrawekanî	[‘damez’, ‘rawe’, ‘kanî’]	[دامەز، راو، ەکانی]	damezraw-ekan-î
BPE	2514/2446	lelayekewe	[’lelay’, ’ek’, ’ewe’]	[لەلای ، ەک ، ەوە]	yekêtî-ye
BPE	2514/2446	gořanêk	[‘gořan’, ‘êk’]	[گۆڕان، تک]	gořan-êk

Table 3: Token and morpheme segmentation examples across Unimorph and BPE tokenizers

under-resourced language contexts.

While recognizing the contributions of traditional unsupervised segmenters such as Morfessor (Creutz and Lagus, 2005), Adaptor Grammars (Johnson et al., 2006) and DPSeg (Dirichlet Process-based Segmenter) (Goldwater et al., 2005) in morphological analysis, this research primarily explores the application these subword tokenizers that are used in modern neural methods. We will extend this comparison to include these traditional segmenters, particularly focusing on their unigram versions which share similarities with the Unigram model used in this study. For future work, we wish to explore the effects of using smaller training datasets with less bias in frequency distribution, build tokenization models based on vocabularies rather than corpora, and train greedy contextual decoding tokenizers (e.g., Uzan et al., 2024).

7 Conclusion

In this study, we have explored the capacity of different tokenization models to segment Soranî Kurdish text into morphologically well-formed subwords. Our findings highlight the differentical effects of pruning and merging on the inductive biases of these models, shedding light on their ability to capture morphological structures. We find that Unigram tokenization leads to the highest quality off-the-shelf morpheme segmentation and find that data augmentation is a less effective strategy than increased vocabulary size in a monoscript context. This research will contribute to the development of more effective NLP tools for low-resource languages with smaller sources and only vocabulary lists, with a focus on morphologically and phonologically motivated analyses.

References

Sina Ahmadi. 2020a. KLPT – Kurdish language processing toolkit. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 72–84, Online. Association for Computational Linguistics.

Sina Ahmadi. 2020b. A tokenization system for the Kurdish language. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 114–127, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Sina Ahmadi, Hossein Hassani, and Kamaladdin Abedi. 2020. A corpus of the Sorani Kurdish folkloric lyrics. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 330–335, Marseille, France. European Language Resources association.

Mohamed Alkaoud and Mairaj Syed. 2020. On the importance of tokenization in Arabic embedding models. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 119–129, Barcelona, Spain (Online). Association for Computational Linguistics.

Tamali Banerjee and Pushpak Bhattacharyya. 2018. Meaningless yet meaningful: Morphology grounded subword-level NMT. In *Proceedings of the Second Workshop on Subword/Character LEvel Models*, pages 55–60, New Orleans. Association for Computational Linguistics.

Yohannes Biadgligne and Kamel Smaili. 2023. Baseline transliteration corpus for improved english-amharic machine translation. *Informatica*, 47(6).

Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4617–4624, Online. Association for Computational Linguistics.

Michael L Chyet and Martin Schwartz. 2003. *Kurdish-English Dictionary*. Yale University Press.

Mathias Creutz and Krista Lagus. 2005. Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.0. *Helsinki University of Technology*.

Kyumars Sheykh Esmaili, Donya Eliassi, Shahin Salavati, Purya Aliabadi, Asrin Mohammadi, Somayeh Yosefi, and Shownem Hakimi. 2013. Building a test collection for sorani kurdish. In *2013 ACS International Conference on Computer Systems and Applications (AICCSA)*, pages 1–7.

Philip Gage. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Leipzig corpora collection. Available online at https://corpora.uni-leipzig.de/.

Sharon Goldwater, Thomas L Griffiths, and Mark Johnson. 2005. Interpolating between types and tokens by estimating power-law generators. In *Advances in neural information processing systems*.

Mark Johnson, Thomas L Griffiths, and Sharon Goldwater. 2006. Adaptor grammars: A framework for specifying compositional nonparametric bayesian models. In *Advances in neural information processing systems*.

Vani Kanjirangat, Tanja Samardžić, Ljiljana Dolamic, and Fabio Rinaldi. 2023. Optimizing the size of subword vocabularies in dialect classification. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 14–30, Dubrovnik, Croatia. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Aso Mahmudi and Hadi Veisi. 2021. Automated grapheme-to-phoneme conversion for central kurdish based on optimality theory. *Computer Speech Language*, 70:101222.

Shervin Malmasi. 2016. Subdialectal differences in Sorani Kurdish. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 89–96, Osaka, Japan. The COLING 2016 Organizing Committee.

Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. 2015. Sarcasm detection on twitter: A behavioral modeling approach. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, page 97–106, New York, NY, USA. Association for Computing Machinery.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Ali Shazal, Aiza Usman, and Nizar Habash. 2020. A unified model for Arabizi detection and transliteration using sequence-to-sequence models. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 167–177, Barcelona, Spain (Online). Association for Computational Linguistics.

Omri Uzan, Craig W. Schmidt, Chris Tanner, and Yuval Pinter. 2024. Greed is all you need: An evaluation of tokenizer inference methods.

Hadi Veisi, Mohammad MohammadAmini, and Hawre Hosseini. 2019. Toward Kurdish language processing: Experiments in collecting and processing the AsoSoft text corpus. *Digital Scholarship in the Humanities*, 35(1):176–193.

8 Appendix

The python version used in this paper is 3.9.6. The Hugging Face tokenizer library version 0.15.2 is used for training BPE and Unigram models. Sentencepiece is trained using version 0.2.0.

Decomposing Fusional Morphemes with Vector Embeddings

Michael Ginn and **Alexis Palmer**
University of Colorado
michael.ginn@colorado.edu and alexis.palmer@colorado.edu

Abstract

Distributional approaches have proven effective in modeling semantics and phonology through *vector embeddings*. We explore whether distributional representations can also effectively model morphological information. We train static vector embeddings over morphological sequences. Then, we explore morpheme categories for *fusional morphemes*, which encode multiple *linguistic dimensions*, and often have close relationships to other morphemes. We study whether the learned vector embeddings align with these linguistic dimensions, finding strong evidence that this is the case. Our work uses two low-resource languages, Uspanteko and Tsez, demonstrating that distributional morphological representations are effective even with limited data.

1 Introduction

Distributional semantics, which models the meanings of words according to the contexts in which they appear (Wittgenstein, 1953), has proven highly successful for language modeling. Generally, this has been achieved through **word embeddings**, which represent words with many-dimensional vectors (Turney and Pantel, 2010; Mikolov et al., 2013b; Levy and Goldberg, 2014b), and capture many linguistic patterns and regularities (Mikolov et al., 2013b; Levy and Goldberg, 2014a).

Linguistic research has suggested that this distributional approach can be effective across all units of language (Haas, 1954). Prior work (Silfverberg et al., 2018; Kolachina and Magyar, 2019) has explored a distributional approach to phonology, finding that embeddings for phonological units can capture predictable linguistic features and natural classes.

We explore whether this approach is also useful for morphology, hypothesizing that many grammatical morphemes can be described primarily by the contexts in which they appear. For example, a first person verbal affix might typically co-occur with first person pronouns, depending on the properties of the language being modeled.

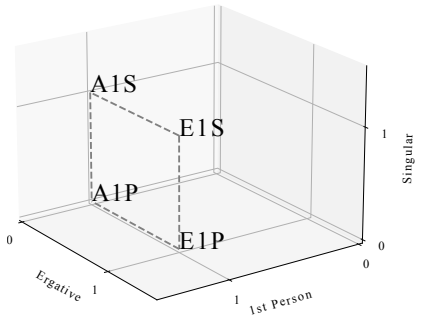

Figure 1: Morpheme glosses in a handcrafted linguistic feature space. Related glosses have predictable vector relationships. A=absolutive case, E=ergative case, P=plural number, S=singular number, 1=first person.

We focus on groups of highly related morphemes, in particular instances of **fusional morphology**. Languages with fusional morphology include single morphemes that encode multiple grammatical features (as opposed to agglutinating morphology, where each morpheme corresponds to a single grammatical function). It is disputed whether languages exist with solely agglutinating or fusional morphological systems; rather, evidence suggests that many languages incorporate both processes (Plank, 1999; Haspelmath, 2009).

We compute morphological embeddings using standard vector embedding algorithms on morphological sequences from two low-resource languages, Uspanteko and Tsez (section 3). We compare these embeddings to handcrafted feature vectors based on the *linguistic dimensions* that make up the morphemes (see Figure 1). We find that there is a consistent correlation between the vector embedding space and this linguistic feature space.

Proceedings of the The 21st SIGMORPHON workshop on Computational Morphology, Phonology, and Phonetics, pages 57–66

2 Data and Languages

Data Format We utilize morpheme sequences from **interlinear glossed text** (IGT) data, a format commonly used in language documentation. An example of Uspanteko IGT is given in item 1.[1]

(1) Ti- j- ya' -tq -a' juntiir
INC- E3S- VT -PL -ENF ADV
They give us everything
(Pixabaj et al., 2007)

The first line records text in the target language. The second line, referred to as the *gloss line*, is a sequence of morphological glosses for each morpheme in the transcription, describing the morphological category and function of each morpheme. Often, stem morphemes may instead by glossed with a translation of the stem, however, in this work we use morphological category glosses as exemplified here (e.g. VT for the transitive verb stem *ya'*). The last line in an IGT example is generally a translation into English or a similarly high-resource language. We utilize only the gloss lines of IGT as morphological category sequences.

We use data from Ginn et al. (2023), which we have formatted in HuggingFace datasets, available online.[2] We use the train splits from Ginn et al. (2023), with 9,774 Uspanteko sentences and 7,116 Tsez sentences.

Languages Uspanteko (usp), or Uspantek, is an endangered Mayan language of Guatamala with around 6,000 speakers (Bennett et al., 2016). The language uses a system of absolutive and ergative affixes which generally attach to verbal stems (Coon, 2016). These affixes are fusional, encoding case (absolutive or ergative), number (singular or plural), and person (first, second, or third-person).

Tsez (ddo), or Dido, is a language in the Nakh-Daghestanian family, with around 14,000 speakers in Daghestan, Russia. Tsez utilizes a highly agglutinating and fusional morphological system, with morphemes often encoding two to five distinct linguistic dimensions. Our data is originally from the Tsez Annotated Corpus Project (Abdulaev et al., 2022; Abdulaev and Abdullaev, 2010).

3 Static Morphological Embeddings

We first investigate whether distributional representations are applicable to morphological sequences—that is, do the contexts that morphemes occur in reflect any meaningful linguistic relationships, and can we capture those relationships with distributional methods? To do this, we train embeddings over sequences of morphological categories from the gloss lines of the IGT from the corpora described in section 2.

We might also have trained embeddings over the morphemes themselves, rather than their glosses/categories. However, our corpora are rather small, and the majority of morphemes occur very rarely, making it difficult to induce meaningful representations. By studying sequences of morpheme categories, we can gain insight into broader morphological patterns, despite limited data.

3.1 Models

Following the approach used in Silfverberg et al. (2018), we consider two different models for learning morphological category embeddings. In all cases, directionality is not considered, so we treat neighboring glosses uniformly, regardless of whether they precede or follow the target gloss.

SVD We compute *positive pointwise mutual information* (PPMI) matrices for each morpheme category in some context window and calculate the *singular value decomposition* (SVD) (Bullinaria and Levy, 2007; Levy and Goldberg, 2014b). We truncate embeddings to some vector length d.

word2vec The word2vec (Mikolov et al., 2013a) model uses a shallow neural network, trained to predict the surrounding words in a sliding window, using the embedding layer as word representations. We use the `gensim` implementation[3] with the default parameters (including negative sampling) and experiment with both the skip-gram and continuous bag-of-words (CBOW) algorithms.

3.2 Experimental Settings

We train separate embedding models over the Uspanteko and Tsez morpheme sequences. For both model types (SVD and word2vec), we train models with vector sizes of 5 to 50 and window sizes of 1 to 10, for a total of 460 distinct runs for each language-model combination. We omit any glosses with fewer than five occurrances.

We believe it is important to report results across hyperparameter combinations, as this is an unsupervised task where it is difficult to tune hyperparameters, and using only a single combination of

[1] A full table of gloss definitions appears in Appendix B.

[2] https://huggingface.co/datasets/lecslab/usp-igt, https://huggingface.co/datasets/lecslab/ddo-igt

[3] https://radimrehurek.com/gensim/

Gloss	Most similar gloss		
	SVD	W2V (CBOW)	W2V (SG)
Uspanteko			
A1P	A2P	A2S	A2S
E1P	A2P	E3	E3
S (noun)	AFI	SREL	SREL
VI	A2P	VT	VT
Tsez			
DEM1.IPL	VOC	DEM2.IPL	DEM2.IPL
DEM2.IISG.OBL	VOC	DEM2.ISG.OBL	DEM2.ISG.OBL
POSS.ESS	COND.IRR	POSS.LAT	LAT
SUPER.ESS	IRR	IN.ESS	CONT.ESS

Table 1: For each gloss embedding, the gloss with the most similar embedding. Here we present a subset of interesting results, full results are in Appendix B.

parameters may produce results which are unrepresentative of the typical performance.

3.3 Results

3.3.1 Related glosses have similar embeddings

First, we investigate whether linguistically-related glosses tend to occur in similar contexts. For each gloss (e.g. A1S), and for every hyperparameter setting, we compute the most similar (distinct) embedding to the gloss's embedding, using cosine similarity. Then, for each gloss we select the most common similar gloss across hyperparameter settings. We highlight a subset of interesting results in Table 1, and report the full results in Appendix B.

We observe differences between the models. The `word2vec` models are far more likely to capture linguistically interesting similarities, while the SVD model does so much less reliably. In the `word2vec` results, closely related glosses, such as VI (intransitive verb stem) and VT (transitive verb stem) tend to be very similar. Both `word2vec` models predict SREL (relational noun) as the most similar gloss to S (noun). Additionally, fusional morpheme glosses such as E1S (ergative first-person singular) tend to be similar to other fusional glosses with the same features, such as E2S (ergative second-person singular). The results for Tsez show similar patterning, with `word2vec` models more closely aligning glosses representing related categories.

3.3.2 Gloss embedding spaces correlate with linguistic feature spaces

Following Silfverberg et al. (2018), we conduct a quantitative measurement in order to understand whether the geometry of the embedding space correlates with a space defined by manually chosen linguistic features. We do not make any assumptions about the magnitude or orientation of embedding vectors; rather, we focus on the cosine similarity scores between embedding vector pairs.

Specifically, we assign vectors to the fusional morphemes in each dataset, using the linguistic dimensions defined in the UniMorph schema (Kirov et al., 2016) as features. Unlike the phonological feature spaces of Silfverberg et al. (2018), it is difficult to decompose all glosses into a single set of linguistic dimensions, as many glosses are completely unrelated. Instead, we focus on the subset of morpheme glosses which share clear features. Each linguistic feature value (e.g. ergative case) is represented as a binary dimension, as in Figure 2. We describe the glosses and linguistic dimensions in detail in Appendix B.

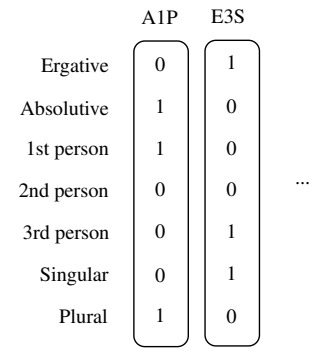

Figure 2: Each morpheme gloss is assigned a handcrafted linguistic feature vector, based on linguistic dimensions from the Unimorph schema. Two examples in Uspanteko are shown here.

For a pair of fusional morpheme glosses, we compute the cosine similarity of the linguistic feature vectors for each gloss. We also compute the cosine similarity for the same glosses using the embedding vectors from the embedding model. We aggregate these similarity measurements across all pairs of glosses that have at least one feature in common. Glosses without any features in common are orthogonal in the linguistic space, hence similar-

ity will be 0. As embedding vectors will generally never have a similarity of 0, we found this added significant noise to the correlation calculation.

Then, we compute the linear correlation coefficient between the linguistic space similarities and the embedding space similarities. As a baseline, we select a random vector in the embedding space for each gloss vector, compute similarities, and calculate the correlation coefficient with the linguistic space similarities. We conduct this process over the hyperparameter combinations described above and report summary results in Table 2 and box plots in Figure 3 and Figure 4.

	Mean / max correlation coefficient r		
	SVD	W2V (CBOW)	W2V (SG)
Uspanteko			
Random	0.05 / 0.49	-0.06 / 0.27	-0.03 / 0.35
True	0.26 / **0.68**	0.19 / 0.42	**0.36** / 0.50
Tsez			
Random	0.02 / 0.10	-0.04 / 0.08	-0.04 / 0.06
True	**0.21** / **0.27**	0.08 / 0.13	0.12 / 0.19

Table 2: Mean / max correlations between linguistic feature space and embedding feature spaces, across hyperparameters.

Findings Broadly, we find that the correlations between the linguistic feature spaces and the vector embedding spaces are greater than the correlations with randomly-selected vector embedding spaces, with the SVD models achieving the highest max correlation across languages. We conduct a paired t-test between the random and true correlation values for each model and language, and find that there is a statistically significant difference in every case

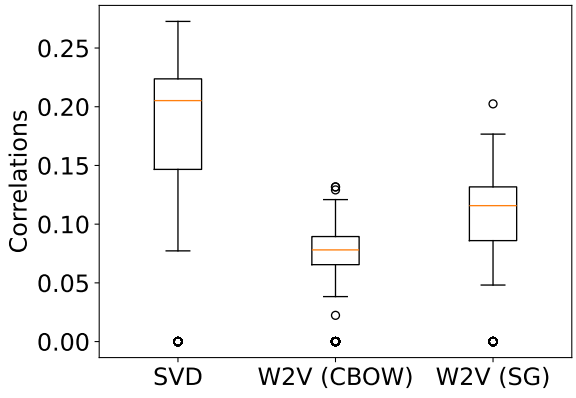

Figure 3: Box plots for Tsez correlation values across hyperparameter values.

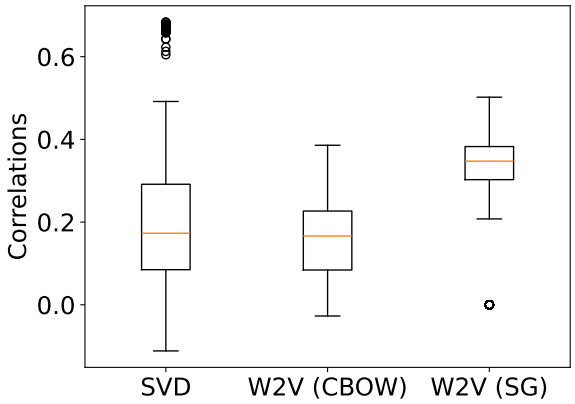

Figure 4: Box plots for Uspanteko correlation values across hyperparameter values.

with $p < 0.001$

The mean correlations are still fairly low—this is likely due in part to the small size of the dataset, but may also indicate that the models are learning relationships between morphemes other than the linguistic dimensions we specify. Future work could investigate these vector spaces more thoroughly to search for novel morphological relationships.

Hyperparameters Not all hyperparameter values perform equally well. We report heatmaps for each model across window size and vector size in Figure 5 and Appendix A. For SVD models, correlation with the linguistic space is maximized with small window sizes (1-2) and decreases significantly with greater window sizes, indicating that the features captured by our linguistic dimensions are generally locally predictable. On the other hand, the `word2vec` models seem to have more consistent performance across window sizes, perhaps indicating that the models are more robust against the noise induced with larger windows. None of the models show significant differences across vector sizes, although the SVD models perform poorly with large windows and very small vector sizes.

4 Related Work

Word embeddings (Turney and Pantel, 2010; Mikolov et al., 2013a,b; Levy and Goldberg, 2014b) have been widely successful in NLP, capturing semantic relationships in many-dimensional vector representations.

Vector embeddings have been applied to phonology, where *phone embeddings* have been used to capture phonetic relationships (Silfverberg et al., 2018; Kolachina and Magyar, 2019; Mayer, 2020).

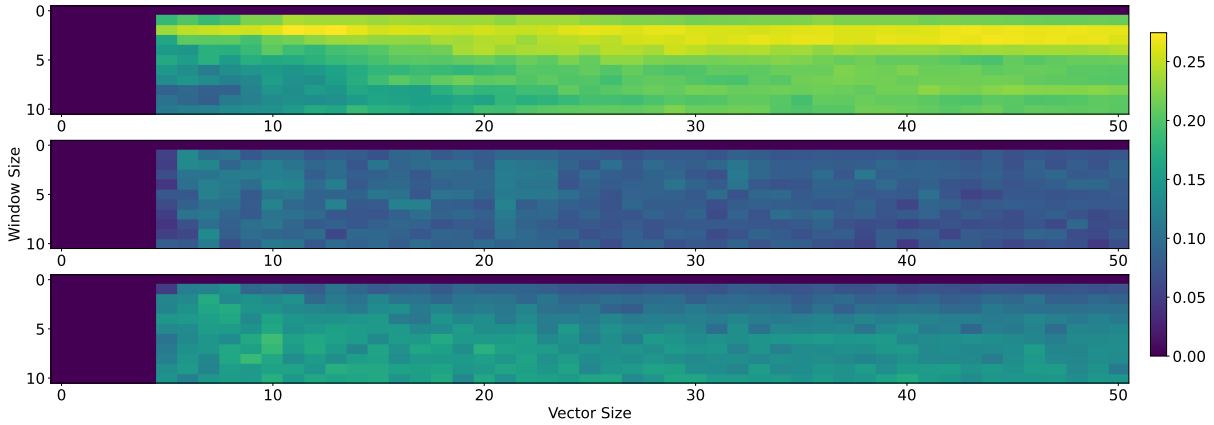

Figure 5: Heatmaps for Tsez of vector space correlation over hyperparameters between the linguistic feature space and the embedding spaces produced by the SVD (top), CBOW (middle), and Skip-gram (bottom) models.

Morphological information has been integrated into word embeddings to improve representations in morphologically-rich languages (Cao and Rei 2016; Edmiston and Stratos 2018; Ataman and Federico 2018; Schwartz et al. 2022, inter alia). To our knowledge, this is the first work that explores a distinct level of morpheme embeddings.

5 Conclusion

We find evidence that distributional vector representations of morpheme categories capture linguistic regularities, even with limited data. Broadly, morphological features such as number, case, and person seem to correlate with the contexts those morphemes appear in. We suggest that distributional morpheme representations are a viable model for morphology, particularly in languages with highly-productive, fusional morphemes.

This research is motivated primarily by linguistic understanding; that is, we are interested in determining whether morpheme contexts have predictable relationships. However, we suggest these findings could be applied in future research to more practical ends. For example, a linguist might use this approach to investigate a hypothesis about the relatedness of certain morphemes, providing for data-driven, large-scale evidence. Alternately, an NLP practioner could use these findings in a task such as morpheme glossing (Ginn et al., 2023) to design models that utilize shared features to make predictions.

6 Limitations

Our research utilizes morphological datasets from two distinct languages. However, considering the linguistic diversity of the world's languages, we expect results may vary across additional languages. In particular, languages without fusional morphology may not show strong linguistic correlations, like we observed in this work.

Acknowledgments

This work utilized the Blanca condo computing resource at the University of Colorado Boulder. Blanca is jointly funded by computing users and the University of Colorado Boulder. Portions of this work were supported by the National Science Foundation under Grant No. 2149404, "CAREER: From One Language to Another". Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

A. K. Abdulaev and I. K. Abdullaev, editors. 2010. *Cezyas folklor/Dido (Tsez) folklore/Didojskij (cezskij) fol ´klor*. "Lotos", Leipzig–Makhachkala.

A.K. Abdulaev, I.K. Abdullaev, André Müller, Evgeniya Zhivotova, and Bernard Comrie. 2022. The Tsez Annotated Corpus Project.

Duygu Ataman and Marcello Federico. 2018. Compositional Representation of Morphologically-Rich Input for Neural Machine Translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 305–311, Melbourne, Australia. Association for Computational Linguistics.

Ryan Bennett, Jessica Coon, and Robert Henderson. 2016. Introduction to Mayan Linguistics. *Lang. Linguistics Compass*, 10:455–468.

John A Bullinaria and Joseph P Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods*, 39:510–526.

Kris Cao and Marek Rei. 2016. A joint model for word embedding and word morphology. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 18–26, Berlin, Germany. Association for Computational Linguistics.

Jessica Coon. 2016. Mayan Morphosyntax: Mayan Morphosyntax. *Language and Linguistics Compass*, 10(10):515–550.

Daniel Edmiston and Karl Stratos. 2018. Compositional morpheme embeddings with affixes as functions and stems as arguments. In *Proceedings of the Workshop on the Relevance of Linguistic Structure in Neural Architectures for NLP*, pages 1–5, Melbourne, Australia. Association for Computational Linguistics.

Michael Ginn, Sarah Moeller, Alexis Palmer, Anna Stacey, Garrett Nicolai, Mans Hulden, and Miikka Silfverberg. 2023. Findings of the SIGMORPHON 2023 shared task on interlinear glossing. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 186–201, Toronto, Canada. Association for Computational Linguistics.

W. Haas. 1954. On defining linguistic units. *Transactions of the Philological Society*, 53(1):54–84.

Martin Haspelmath. 2009. *An Empirical Test of the Agglutination Hypothesis*, pages 13–29. Springer Netherlands, Dordrecht.

Christo Kirov, John Sylak-Glassman, Roger Que, and David Yarowsky. 2016. Very-large scale parsing and normalization of Wiktionary morphological paradigms. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3121–3126, Portorož, Slovenia. European Language Resources Association (ELRA).

Sudheer Kolachina and Lilla Magyar. 2019. What do phone embeddings learn about phonology? In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 160–169.

Omer Levy and Yoav Goldberg. 2014a. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 171–180, Ann Arbor, Michigan. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014b. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Connor Mayer. 2020. An algorithm for learning phonological classes from distributional similarity. *Phonology*, 37(1):91–131.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013a. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Telma Can Pixabaj, Miguel Angel Vicente Méndez, Maria Vicente Méndez, and Oswaldo Ajcot Damián. 2007. Text collections in four mayan languages. *Archived in The Archive of the Indigenous Languages of Latin America*.

Frans Plank. 1999. Split morphology: How agglutination and flexion mix. *Linguistic Typology*, 3:279–340.

Lane Schwartz, Coleman Haley, and Francis Tyers. 2022. How to encode arbitrarily complex morphology in word embeddings, no corpus needed. In *Proceedings of the first workshop on NLP applications to field linguistics*, pages 64–76, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Miikka Silfverberg, Lingshuang Jack Mao, and Mans Hulden. 2018. Sound analogies with phoneme embeddings. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2018*, pages 136–144.

Peter D Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*.

Ludwig Wittgenstein. 1953. *Philosophical Investigations*. Blackwell, Oxford.

A Uspanteko Heatmap

We provide the correlation heatmap for Uspanteko, similar to the Tsez figure provided in the main paper in Figure 6.

B Glosses

We report a complete list of the glosses in each language in Table 3 and Table 4.

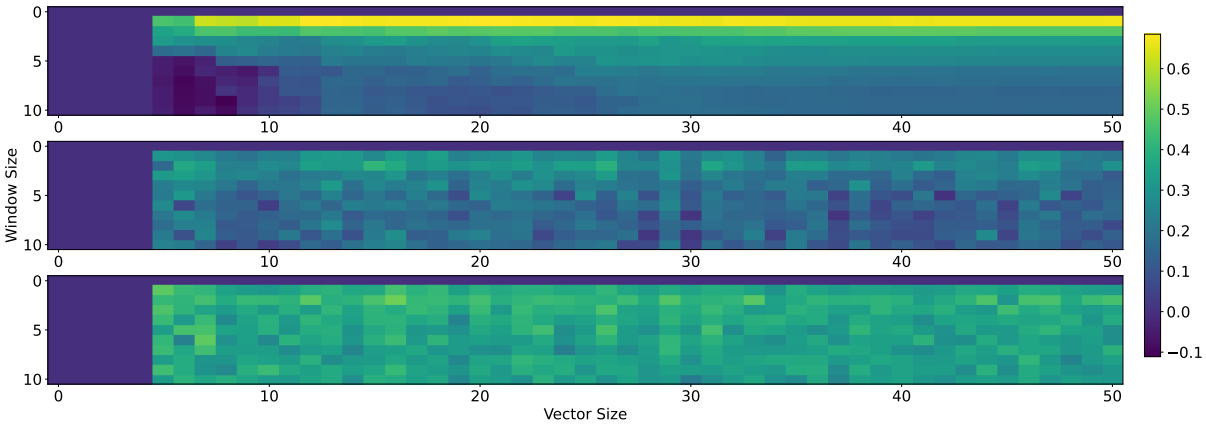

Figure 6: Heatmaps for Uspanteko of vector space correlation over hyperparameters between the linguistic feature space and the embedding spaces produced by the SVD (top), CBOW (middle), and Skip-gram (bottom) models.

Gloss	Label	Count	Features	Most similar gloss SVD	CBOW	SG
A1P	Absolutive 1P Pl	110	Abs., 1st, Pl.	A2P	A2S	A2S
A1S	Absolutive 1P Sing	347	Abs., 1st, Sing.	REC	A2S	A2S
A2S	Absolutive 2P Sing	127	Abs., 2nd, Sing.	DIM	A1P	A1P
ADJ	Adjective Stem	1017		APLI	NUM	ITS
ADV	Adverb Stem	5830		APLI	PART	PART
AFE	Affective	116		A2P	PREP	PREP
AFI	Positive	208		E3P	PART	DEM
AGT	Agentive	100		A2P	E2	E2
AP	Antipassive	339		A2P	E2S	E2S
ART	Article	973		INT	NUM	NUM
CAU	Causative	19		PRG	GNT	RFX
CLAS	Classifier	155		REC	NOM	NOM
COM	Completive	2304		NOM	INC	PP
COND	Conditional	59		REC	IMP	PRG
CONJ	Conjunction	1152		A2P	VOC	VOC
DEM	Dem.	2116		APLI	AFI	AFI
DIM	Diminutive	797		A2S	ART	NUM
DIR	Directional	687		A2P	PAS	PAS
E1P	Ergative 1P Pl	1370	Erg., 1st, Pl.	A2P	E3	E3S
E1S	Ergative 1P Sing	709	Erg., 1st, Sing.	NOM	E2S	E2S
E2	Ergative 2P	16	Erg., 2nd	NOM	INS	RFX
E2P	Ergative 2P Pl	16	Erg., 2nd, Pl.	ART	INS	E2
E2S	Ergative 2P Sing	564	Erg., 2nd, Sing.	A2P	AP	AP
E3	Ergative 3P	385	Erg., 3rd	A2P	E3S	E3S
E3P	Ergative 3P Pl	32	Erg., 3rd, Pl.	NOM	E2P	E2P
E3S	Ergative 3P Sing	3118	Erg., 3rd, Sing.	NOM	E3	E3
ENF	Emphasis	1464		A2P	A1P	IMP
EXS	Existential	661		A1P	NUM	NUM
GNT	Demonym	20		TRN	INS	RFX
IMP	Imperative	67		EXS	COND	COND
INC	Incompletive	2742		NOM	COM	SC
INS	Instrumental	37		A2P	GNT	E2P
INT	Interrogative	343		ART	NEG	NEG
ITR	Intransitive	73		A2P	E2	RFX
ITS	Intensifier	244		GNT	AFI	ADJ
MED	Measure	66		A2P	POS	AGT
MOV	Auxiliary	141		REC	AGT	AGT
NEG	Negative	1130		REC	INT	INT
NOM	Proper Name	167		PAS	CLAS	CLAS
NUM	Numeral	1029		APLI	ART	MED
PART	Particle	3153		A2P	ADV	ADV
PAS	Passive	276		A2P	DIR	E3P
PL	Pl	2094		DEM	PREP	PREP
POS	Positional	83		E2P	MED	GNT
PP	Perfect Participle	127		REC	AGT	AGT
PREP	Preposition Stem	1605		A2P	PL	AFE
PRG	Progressive	42		CAU	GNT	TRN
PRON	Pronoun	1674		REC	INT	A2S
RFX	Reflexive	8		INT	GNT	TRN
S	Noun Stem	6962		AFI	SREL	SREL
SAB	Abstract Noun Stem	158		CONJ	MED	INS
SC	Category Suffix	1018		A2P	ENF	SV
SREL	Relative Noun	1890		TRN	S	S
SV	Verbal Noun Stem	88		E1	INS	INS
TAM	Tense-Aspect-Mood	128		APLI	SV	SV
TOP	Proper Noun Stem	108		A2P	MED	GNT
TRN	Applicative	7		TOP	GNT	RFX
VI	Intransitive Verb Stem	3125		A2P	VT	VT
VOC	Vocative	750		A2P	CONJ	CONJ
VT	Transitive Verb Stem	5024		NOM	VI	VI

Table 3: All of the glosses in Uspanteko, along with a description, the total number of occurrances, and a list of positive features in the linguistic vector representations.

Gloss	Label	Most similar gloss SVD	CBOW	SG
AD.ABL	Position At, Ablative	PST.UNW	APUD.ABL	AD.VERS
AD.ESS	Position At, Essive	APUD.VERS.DIST	SUB.ABL	SUB.ABL
AD.LAT	Position At, Lative	COND.IRR	IN.ESS	IN.ESS
AD.VERS	Position At, Versative	POSS.ESS.DIST	SUPER.VERS	CONT.VERS
AD.VERS.DIST	Position At, Versative, Distal	PROHIB	APUD.ABL	CONT.ABL.DIST
ANT.CVB	Anterior, Converb	COND.IRR	IMM.ANT.CVB	IMM.ANT.CVB
APUD.ABL	Pos. Near, Ablative	DEM2.IIPL	INT	AD.VERS.DIST
APUD.ESS	Pos. Near, Essive	PST.UNW	APUD.VERS	APUD.LAT
APUD.LAT	Pos. Near, Lative	APUD.ABL.DIST	APUD.VERS	APUD.VERS
APUD.VERS	Pos. Near, Versative	PST.UNW	APUD.LAT	APUD.LAT
APUD.VERS.DIST	Pos. Near, Versative, Distal	SUPER.LAT.DIST	DEM3.SG	LOC.ORIG
ATTR	Attributive	NEG.PRSPRT.OBL	ATTR.OBL	RES.PRT.OBL
ATTR.OBL	Attributive, Oblique	SUPER.ESS.DIST	ATTR	GEN2
CNC.CVB	Concessive, Converb	DEM2.IIPL	COND	COND
CND	Conditional	SUPER.LAT.DIST	CONT.ABL.DIST	IN.LAT.DIST
CND.CVB	Conditional, Converb	DIST	PRS.PRT	COND
CND.CVB.IRR	Conditional, Converb, Irrealis	SUPER.LAT.DIST	COND	DEM3.SG
COND	Conditional	SUPER.LAT.DIST	DEM3.SG	NEG.PRS.PRT
COND.IRR	Conditional, Irrealis	SUPER.LAT.DIST	DUB	INDEF
CONT.ABL	Pos. Among, Ablative	GER.PURP	IN.ESS	GEN1
CONT.ABL.DIST	Pos. Among, Ablative, Distal	EQU1	APUD.ABL	IN.LAT.DIST
CONT.ESS	Pos. Among, Essive	SUPER.ESS.DIST	GEN1	POSS.ABL
CONT.LAT	Pos. Among, Lative	NEG.PRS.PRT.OBL	SUB.ESS	SUB.ESS
CONT.VERS	Pos. Among, Versative	NEG.PRSPRT	AD.VERS	IN.ABL
CONT.VERS.DIST	Pos. Among, Versative, Distal	SUPER.LAT.DIST	IN.ESS.DIST	SUPER.ABL.DIST
CSL.CVB	Causal, Converb	IN.VERS.DIST	INF	NEG.PST.UNW
DEF	Definite	SUPER.ESS.DIST	AD.LAT	SUB.LAT
DEM1.IIPL	C1 Dem. 2nd N Pl	PST.UNW	POSS.VERS	DEM1.IIPL.OBL
DEM1.IIPL.OBL	C1 Dem. 2nd N Pl, Oblique	VOC	DEM2.IPL.OBL	DEM1.IIPL
DEM1.IISG.OBL	C1 Dem. 2nd N Sing, Oblique	SUPER.LAT.DIST	DEM2.ISG.OBL	DEM3.IISG.OBL
DEM1.IPL	C1 Dem. 1st N Pl	VOC	DEM2.IPL	DEM2.IPL
DEM1.IPL.OBL	C1 Dem. 1st N Pl, Oblique	RES.PRT.OBL	DEM2.IPL.OBL	DEM1.IPL
DEM1.ISG.OBL	C1 Dem. 1st N Sing, Oblique	NEG.PST.UNW	DEM2.ISG.OBL	DEM2.IISG.OBL
DEM1.SG	C1 Dem. Sing	APUD.VERS.DIST	II	DEM4.SG
DEM2.IIPL.OBL	C2 Dem. 2nd N Pl, Oblique	DEM1.IISG	DEM3.IISG.OBL	DEM3.IISG.OBL
DEM2.IISG	C2 Dem. 2nd N Sing	APUD.ABL.DIST	PROHIB	DEM1.SG
DEM2.IISG.OBL	C2 Dem. 2nd N Sing, Oblique	VOC	DEM2.ISG.OBL	DEM2.ISG.OBL
DEM2.IPL	C2 Dem. 1st N Pl	CND.CVB	DEM1.IPL	DEM1.IPL
DEM2.IPL.OBL	C2 Dem. 1st N Pl, Oblique	NEG.PRS.PRT	DEM1.IIPL.OBL	DEM2.IPL
DEM2.ISG	C2 Dem. 1st N Sing	LNK	IN.LAT	IN.ESS.DIST
DEM2.ISG.OBL	C2 Dem. 1st N Sing, Oblique	NEG.PST.UNW	DEM1.ISG.OBL	DEM2.IISG.OBL
DEM2.PL	C2 Dem. 2nd N Pl	DEM3.IPL	POSS.VERS	CONT.VERS.DIST
DEM3.IISG.OBL	C3 Dem. 2nd N Sing, Oblique	SUPER.LAT.DIST	DEM2.IIPL.OBL	INTS
DEM3.SG	C3 Dem. Sing	SUPER.LAT.DIST	COND	COND.IRR
DEM4.IISG.OBL	C4 Dem. 2nd N Sing, Oblique	SUPER.LAT.DIST	DEM1.IIPL.OBL	LCV
DEM4.ISG.OBL	C4 Dem. 1st N Sing, Oblique	NEG.PST.UNW	DEM3.IISG.OBL	CONT.ABL.DIST
DEM4.SG	C4 Dem., Sing	SUPER.LAT.DIST	DEM4.ISG.OBL	DEM4.ISG.OBL
FUT.CVB	Future, Converb	SUB.ESS.DIST	NEG.FUT.DEF	NEG.PRS.PRT
FUT.DEF	Future, Definite	LNK	NEG.FUT	NEG.FUT
I.PL	1st Noun, Plural	CND.CVB	DEM2.IPL	DEM2.IPL
II	2nd Noun	LNK	DEM1.SG	DEM2.IISG
II.PL	2nd Noun, Plural	CND.CVB	IV.PL	DEM1.IIPL
III	3rd Noun	LNK	DEM2.ISG	DEM1.SG
III.PL	3rd Noun, Plural	SUB.ESS.DIST	II.PL	DEM1.IIPL
IMM.ANT.CVB	Immediate, Anterior, Converb	NEG.PST.UNW	POST.CVB	POST.CVB
IN.ABL	Position In, Ablative	NEG.PST.UNW	IN.ALL	CONT.VERS
IN.ABL.DIST	Position In, Ablative, Distal	CND.CVB	IN.ESS.DIST	CONT.VERS
IN.ALL	Position In, Allative	CND.CVB	IN.LAT	IN.VERS.DIST
IN.ESS	Position In, Essive	POSS.ESS.DIST	AD.LAT	AD.LAT
IN.ESS.DIST	Position In, Essive, Distal	POSS.ESS.DIST	APUD.ABL	CONT.ABL.DIST

Table 4: All of the glosses in Tsez, along with a description, the total number of occurrances, and a list of positive features in the linguistic vector representations. C# Dem.=class of demonstratives. The I, II, etc., morphemes indicate the four noun classes of Tsez.

Gloss	Label	Most similar gloss		
		SVD	CBOW	SG
IN.LAT	Position In, Lative	CND.CVB	IN.ALL	IN.ABL.DIST
IN.LAT.DIST	Position In, Lative, Distal	SUPER.LAT.DIST	IN.ESS.DIST	CND
IN.VERS	Position In, Versative	OSS.ESS.DIST	CONT.LAT	AD.VERS
IN.VERS.DIST	Position In, Versative, Distal	SEQ	IN.ESS.DIST	IN.ESS.DIST
INT	Interrogative	APUD.VERS.DIST	APUD.ABL	IN.LAT.DIST
IPFV.CVB	Imperfective, Converb	POSS.ESS.DIST	TERM	TERM
IV	4th Noun	SUPER.LAT.DIST	III	NMLZ
IV.PL	4th Noun, Plural	POSS.ABL.DIST	II.PL	II.PL
LAT	Lative	PST.UNW	POSS.ESS	POSS.ESS
LCV	Locative	GER.PURP	LCV.CVB	POSS.VERS
LCV.CVB	Locative, Converb	PFV.CVB.INT	LCV	LCV
LOC.ORIG	Locative, Origin	GER.PURP	CONT.VERS.DIST	POSS.ABL.DIST
NEG	Negative	SUB.ESS.DIST	Q	Q
NEG.FUT	Negative, Future	SUB.VERS	FUT.DEF	FUT.DEF
NEG.FUT.CVB	Negative, Future, Converb	PST.UNW	COND.IRR	NEG.FUT
NEG.FUT.DEF	Negative, Future, Definite	SUPER.LAT.DIST	PST.WIT.Q	NEG.PRS.PRT
NEG.PRS.PRT	Negative, Present Participle	SUPER.LAT.DIST	NEG.PST.UNW	NEG.PST.UNW
NEG.PRS.PRT.OBL	Neg., Pres. Part., Oblique	APUD.ABL.DIST	CONT.VERS.DIST	POSS.ABL.DIST
NEG.PST.CVB	Negative, Past, Converb	SUB.ESS.DIST	TERM	NEG.PST.UNW
NEG.PST.UNW	Neg., Past, Unwitnessed	DEM2.ISG.OBL	POT	NEG.PRS.PRT
NEG.PST.WIT	Neg., Past, Witnessed	NEG.PST.UNW	Q	PST.WIT.INT
PCT.CVB	Perfective, Converb	IN.VERS	DEM3.SG	POSS.ABL.DIST
PFV.CVB	Perfective, Converb	VOC	EMPH	IN.VERS.DIST
PL	Plural	SUB.ESS.DIST	DEM1.IPL	DEM1.IIPL
POSS.ABL	Position Vertical, Ablative	PST.UNW	APUD.LAT	APUD.LAT
POSS.ABL.DIST	Pos. Vert., Ablative, Distal	SUPER.LAT.DIST	INTS	LOC.ORIG
POSS.ESS	Position Vertical, Essive	APUD.ABL.DIST	POSS.LAT	LAT
POSS.LAT	Position Vertical, Lative	POSS.ESS.DIST	POSS.ESS	GEN1
POSS.VERS	Position Vertical, Versative	COND.IRR	DEM2.PL	AD.VERS.DIST
POST.CVB	Posterior, Converb	LNK	IMM.ANT.CVB	IMM.ANT.CVB
PRS	Present	SUPER.LAT.DIST	FUT.DEF	PST.WIT.Q
PRS.PRT	Present Participle	NEG.PST.UNW	NEG.FUT	NEG.FUT
PRS.PRT.OBL	Present Participle, Oblique	POSS.ESS.DIST	DEM2.IIPL.OBL	DEM4.ISG.OBL
PST.PRT	Past, Participle	DEF1.IISG	ATTR	ATTR
PST.UNW	Past, Unwitnessed	POSS.ESS.DIST	ANT.CVB	ANT.CVB
PST.WIT	Past, Witnessed	PST.UNW	IMPR	NEG.PST.WIT
PST.WIT.INT	Past, Witnessed, Interr.	NEG.PST.UNW	NEG.PST.WIT	NEG.PST.WIT
PST.WIT.Q	Past, Witnessed, Question	IRR	NEG.FUT.DEF	DEM3.IISG.OBL
PURP.CVB	Purposive, Converb	ATTR.OBL	COND	PCT.CVB
Q	Question	AD.ABL.DIST	NEG.PST.WIT	NEG.PST.WIT
RES.PRT	Resultative Participle	SUPER.LAT.DIST	INF	PST.WIT.Q
RES.PRT.OBL	Res. Part., Oblique	LHUN	DEM3.SG	POSS.ABL.DIST
SIM.CVB	Simultaneous Converb	CND.CVB	IMM.ANT.CVB	ANT.CVB
SUB.ABL	Position Under, Ablative	POSS.ESS.DIST	AD.ESS	AD.ESS
SUB.ESS	Position Under, Essive	GER.PURP	CONT.LAT	APUD.ABL
SUB.LAT	Position Under, Lative	SUPER.ESS.DIST	APUD.ABL	CONT.ABL.DIST
SUPER.ABL	Position Under, Ablative	IN.LAT.DIST	CONT.ESS	CONT.ESS
SUPER.ABL.DIST	Pos. Under, Ablative, Distal	NEG.PRSPRT.OBL	INT	POSS.ABL.DIST
SUPER.ESS	Position Above, Essive	IRR	IN.ESS	CONT.ESS
SUPER.LAT	Position Above, Lative	POSS.ESS.DIST	IN.ABL	LCV.CVB
SUPER.VERS	Position Above, Versative	IRR	AD.VERS	IN.ESS.DIST
SUPER.VERS.DIST	Pos. Above, Versative, Distal	GER.PURP	APUD.ABL	APUD.VERS.DIST

Table 5: Tsez glosses (cont.)

Acoustic barycenters as exemplar production targets

Frédéric "Fred" Mailhot
Dialpad, Inc.
fred.mailhot@dialpad.com

Cassandra L. Jacobs
Department of Linguistics
University at Buffalo
cxjacobs@buffalo.edu

Abstract

We present a solution to the problem of exemplar-based language production from variable-duration tokens, leveraging algorithms from the domain of time-series clustering and classification. Our model stores and outputs tokens of phonetically rich and temporally variable representations of recorded speech. We show qualitatively and quantitatively that model outputs retain essential acoustic/phonetic characteristics despite the noise introduced by averaging, and also demonstrate the effects of similarity and indexical information as constraints on exemplar cloud selection.

1 Introduction

We present here an exemplar production model that implements solutions to the challenges of measuring between-exemplar distance (i.e. alignment) and fostering phonetic generalization over speech tokens of variable duration (Pierrehumbert, 2002; Kirchner et al., 2010).[1] Our model, MNEMORPHON, makes use of algorithms for alignment and averaging the domain of time-series clustering and classification. We show qualitatively by direct inspection of model outputs and quantitatively via statistical classification that MNEMORPHON's outputs retain essential acoustic/phonetic characteristics, despite noise introduced by averaging, and also demonstrate the effects constraining exemplar cloud composition by means of similarity weighting and indexical information.

We begin with an overview of exemplar-based approaches to phonetics and phonology, highlighting the core production challenges of temporal variability and generalization. We then introduce MNEMORPHON and present our experiments and results, and finish with some discussion of planned work and future directions.[2]

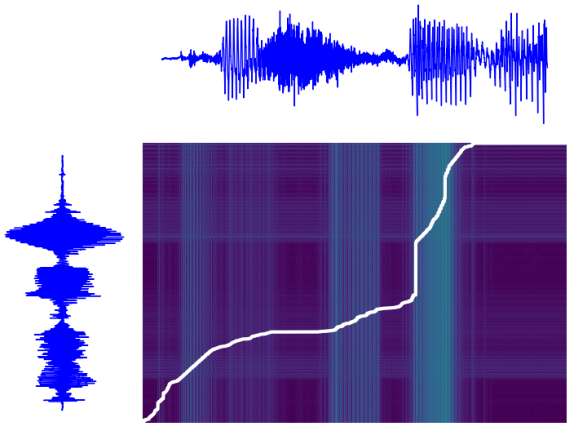

Figure 1: Dynamic time warping alignment of waveforms of two tokens of the Turkish word *kuşları* (*"birds"*), highlighting temporal variability.

2 Exemplar-based phonetics and phonology

Exemplar-based theories of categorization propose that humans classify percepts by direct comparison to memorized exemplars of previous experiences (Semon, 1923; Medin and Schaffer, 1978; Hintzman, 1986; Nosofsky, 1986), whereas linguistic theories have traditionally been couched in terms of symbolic categories that abstract away from details of usage and experience. When experiments in speech perception suggested that human word recognition is facilitated by fine details of remembered experiential episodes, e.g. speakers' voices (Goldinger, 1996, 1998), phoneticians began to explore the possibilities of memory-based approaches. Johnson (1997a,b) presented a pair of exemplar models of phonetic perception that

[1] Here and below, "length", "duration", "variability", etc. specifically refer to *temporal extent*, rather than e.g. number of phones/segments. Where we discuss discrete sequences, it is assumed that sequence coordinates represent a fixed and constant temporal duration.

[2] Code for the model, experiments, and evaluations described below will be made available at https://github.com/calicolab/mnemorphon.

Proceedings of the The 21st SIGMORPHON workshop on Computational Morphology, Phonology, and Phonetics, pages 67–76

provided elegant and novel accounts of speaker normalization and speech segmentation. Soon after Pierrehumbert (2001) published the first implemented model of exemplar-based phonological *production*, in the context of a production-perception feedback-loop model of sound change.

These initial investigations ushered in a flurry of subsequent research in exemplar-theoretic phonetics and phonology in areas as diverse as sound change, categorical emergence and entrenchment, sociophonetic variation, frequency effects in productivity, the status of abstract phonetic categories, and the induction of morphophonological alternations (Bybee, 2001; Pierrehumbert, 2001; Hawkins, 2003; Wedel, 2006; Gahl and Yu, 2006; Johnson, 2006; Ettlinger, 2007; Kirchner et al., 2010; Mailhot, 2010a).

Goldrick and Cole (2023) provide a recent overview of the theoretical and empirical successes, along with some outstanding potential challenges, of exemplar-based approaches to production. The core theoretical challenges faced by exemplar-based models of production are handling input variability, particularly with respect to temporal variation, and the need for a mechanism for robust generalization from prior experiences. Below we discuss the first of these, showing how it can be surmounted with a 50 year old approach to speech recognition, and later we address the latter, introducing a 21st century algorithm for averaging time series.

2.1 The problem of temporal variability

It is well-known that distinct utterances of human speech[3] categories such as words can vary significantly in duration, both within and across speakers (see e.g. Figure 1). This temporal variability is one of the core challenges for any exemplar model. These models typically compute a distance or similarity function over exemplars; we therefore require a means of computing such a measure that is robust to length-wise variation. Fortunately, such an algorithm already exists and is well-known in the speech recognition and time series analysis literatures.

Dynamic time warping (DTW) (Vintsyuk, 1968; Sakoe and Chiba, 1978; Mueen and Keogh, 2016, for a recent overview) is an algorithm for computing a distance measure between sequences of potentially differing lengths. Given a pair of sequences X, Y with *coordinates*[4] $[x_0, ..., x_n], [y_0, ..., y_m]$ embedded in a shared parametric space D^k and a distance function $d(x_i, y_j)$, DTW finds the best alignment between X and Y via the following optimization:

$$DTW(X,Y) = \min_{\pi} \sqrt{\sum_{(i,j)\in\pi} d(x_i, y_j)^2} \quad (1)$$

Here π is an alignment or *warping path* between X and Y; a sequence of pairs $((i_1, j_1), ..., (i_k, j_k))$ each of whose elements respectively indexes positions in X and Y, with the following properties: (i) $\pi_1 = (1, 1)$ and $\pi_k = (n, m)$, that is, the start and end of X and Y are aligned, (ii) π increases monotonically in i and j, and (iii) each $i \in [1, ..., n]$ and $j \in [1, ..., m]$ appears at least once in π. The *DTW distance* between X and Y is the minimized sum of coordinate-wise distances over all possible alignments.

We note here that we are not the first to realize that DTW provides a solution to the problem of temporal alignment in exemplar production; Kirchner et al.'s (2010) PEBLS incorporates it in modeling a phonological generalization on toy data using speech tokens. Their approaches requires ad-hoc modifications to the DTW algorithm, along with an additional hierarchical clustering step to mitigate the problem of spurious generalizations (see Appendix A for a more detailed overview). Below we examine a more principled approach to the latter problem.

2.2 Generalizing production from exemplar knowledge

As alluded to above, a remaining challenge for production exemplar models is accounting for the human capacity to "go beyond the data" and generalize over prior experiences, a hallmark of cognition. In exemplar models of perception/comprehension, a distance measure and simple nearest-neighbour search are sufficient to enable generalization; given an input form, the listener finds the previously stored form that is closest to it in the representational space, and assigns that form's category to the input form, then stores them together.

[3] We focus on speech here and below, but believe the approach developed here applies, *mutatis mutandis*, to signed languages as well.

[4] We borrow this term from Petitjean et al. (2011) for individual (possibly multidimensional) elements of sequences and use it throughout.

In production, the speaker has a given category and must produce an output for it. The simplest means of doing so is to select a previously-memorized token from within that category and directly produce it. This method effectively turns the model into a look-up table, making it in-principle incapable of generalizing beyond the input to which it has been exposed (consider whether this approach could handle e.g. a "wug" test Berko Gleason, 1958). We turn now to one means of surmounting this obstacle.

Pierrehumbert (2001) presents a model of phonological production that implements generalization via a simple but ingenious method of exemplar composition. The model's exemplars are points in $(F1, F2, F3)$ formant space, representing vowel steady-state measurements, paired with vowel category labels. For a given vowel category $\mathcal{C}$, generation of an output exemplar c_{out} proceeds in three steps: (i) a single *seed* token c_{in} is randomly selected from all stored exemplars associated with $\mathcal{C}$, (ii) an analogical set or *exemplar cloud* — C_{in} is constructed by considering all exemplars within a fixed Euclidean distance of c_{in} in formant space, and finally (iii) an output token c_{out} is produced by computing a *similarity-weighted* average of the exemplar cloud, with similarity computed as an inverse exponential function of distance.

Many phonetic and phonological insights have been derived from exemplar models that take inspiration from this approach, averaging over point-like data in low-dimensional spaces (e.g. Wedel, 2006; Ettlinger and Johnson, 2010, *inter alia*). This approach can be straightforwardly extended to handle parametric spaces of higher dimensionality e.g. encoding richer acoustic information with spectral frames, or sociophonetic context such as interlocutor identity, etc. However, it is unclear how it might be extended to incorporate the *dynamic* nature of human language, which unfolds in time and cannot be reduced to point measures. That is, the problem of straightforwardly accommodating the temporal variability and generalization of human speech in implemented production models remains underexplored.[5]

[5]To our knowledge Kirchner et al. (2010)is the only extant model to address it to date.

3 MNEMORPHON: A bit of progress in exemplar-based production[6]

Any implemented exemplar model must minimally include tokens of some primitive linguistic unit encoded in a suitable representational format, associated category labels, and a means of computing analogically relevant similarity between exemplars (Johnson, 2007). For Pierrehumbert's model discussed above, these are segments (specifically vowels), the space defined by tuples of the first three formants, and inverse Euclidean distance in formant space. For MNEMORPHON these architectural parameters are as follows:

- **Units:** tokens are complete words, with no representation of sub-lexical linguistic categories (syllables, segments, etc.)
- **Representation:** exemplars are encoded as mel-scaled spectrograms (Deng and O'Shaughnessy, 2003)
- **Categories:** each exemplar is associated to a discrete "lexical" label encoded as a pseudo-phonemic character string for mnemonic convenience, roughly corresponding to a word meaning
- **Similarity:** similarity between tokens is computed as an inverse function of DTW distance (see below for details)

Our general task can now be framed as follows: given a seed exemplar and a cloud of tokens of possibly varying lengths from a given category, we seek a procedure by which we can generate an output exemplar as an "average" of the cloud.

As it happens, exactly computing the sample mean of a set of sequences with potentially differing lengths corresponds to solving the problem of *multiple sequence alignment*, which is known to be computationally intractable (Elias, 2006). Notwithstanding this, there are tractable approximation methods that are theoretically justifiable and empirically suitable; Petitjean et al. (2011) introduce one such approach, *DTW barycenter averaging* (DBA).

3.1 Computing averages of variable-length sequences

DBA is an algorithm that takes as input a set of sequences and iteratively converges to an average sequence that is locally optimal, in the sense of

[6]With apologies to Goodman (2001).

Algorithm 1 DBA (adapted from Petitjean et al., 2011)

Require: $\mathcal{S}$ the sequences to average
Require: $\hat{s} = [\hat{s}_1 ... \hat{s}_k]$ initial barycenter
  converged $\leftarrow$ False
  assocTbl $\leftarrow$ table of length k
  while converged $\neq$ True **do**
    for all $s \in \mathcal{S}$ **do**
      $\pi \leftarrow \text{DTW}(\hat{s}, s)$
      for all $(i, j) \in \pi$ **do**
        assocTab[i] $\leftarrow$ assocTab[i] $\cup$ s_j
    for all $\hat{s}_i \in \hat{s}$ **do**
      $\hat{s}_i \leftarrow$ BARYCENTER(assocTab(i))
    CHECKCONVERGENCE
  return $\hat{s}$

minimizing a quantity analogous to *inertia* in k-means clustering (i.e. "within-cluster variance" MacQueen, 1967):

$$\hat{s}^* = \min_{\hat{s}} \sum_{s_i \in \mathcal{S}} DTW(\hat{s}, s_i)^2 \qquad (2)$$

The sequence $\hat{s}^*$ is called the *barycenter* of the set of sequences $\mathcal{S}$, by analogy with the use of that term for *center of mass*, a dynamical physical points which need not equal or intersect with any of the points it averages over.

Given a set of sequences $\mathcal{S}$ and an initial "best-guess" barycenter $\hat{s}$ (typically randomly generated or sampled directly from $\mathcal{S}$), DBA iterates over two phases (see Algorithm 1):

- **Align:** compute DTW alignments for $\hat{s}$ and each $s \in \mathcal{S}$, and for each coordinate $\hat{s}_i$ of the barycenter, store the set of all coordinates it was aligned with for each s
- **Update:** update each coordinate $\hat{s}_i$ of $\hat{s}$ to be the barycenter of its associated coordinates found in the alignment phase

The algorithm halts after a predetermined number of iterations, or when the difference in inertia across iterations falls below a preset convergence threshold. At each iteration, the update either moves the barycenter's coordinates to be closer to their aligned cloud elements, or else a lower-cost DTW alignment is found. In either case, the inertia stays the same or decreases, hence DBA is guaranteed to converge.

Algorithm 2 MNEMORPHON output generation

Require: Λ a lexicon of categories and associated exemplars
Require: $\lambda \in \Lambda$ the category for which MNEMORPHON must generate an output
  $S_{in} \leftarrow$ GETALLEXEMPLARS(λ)
  $s_i \leftarrow RandomSelectOne(C_{in})$ ▷ the seed
  $C_{in} \leftarrow$ CONSTRUCTCLOUD(S_{in})
  $\hat{s}^* \leftarrow$ DBA(C_{in}, s_i)) **return** $\hat{s}^*$

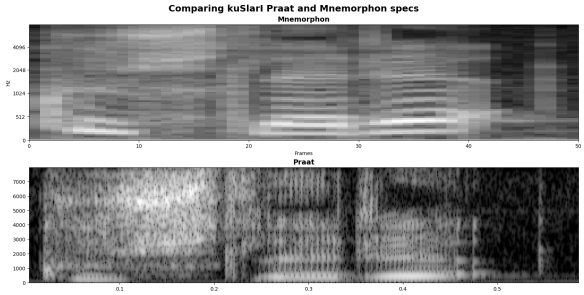

Figure 2: Spectrograms of a token of *kuşları* ("birds"), as created with our parameters versus Praat's default values.

3.2 Generating outputs

With its representations and averaging procedure in place, MNEMORPHON's basic algorithm for exemplar output generation is straightforward (see Algorithm 2):

1. Given a set of stored *(exemplar, category)* pairs, and a target output production category
2. select a *seed* exemplar associated with the target category
3. construct an analogical set or *cloud* from the remaining exemplars in the target category
4. output the mean of the cloud, computed via DBA[7]

We leave the cloud construction step in 3 unspecified here; Pierrehumbert uses a fixed-radius neighbourhood of the seed, but alternatives are possible, e.g. a fixed number of seed neighbours. Below we partially address this question; ultimately it is a model parameter to be tuned empirically.

4 Data

For the experiments described below our raw data set is an audio corpus of Turkish speech, consisting of microphone recordings (16KHz sample rate)

[7] MNEMORPHON uses the implementation of DBA available in the `tslearn` Python package (Tavenard et al., 2020).

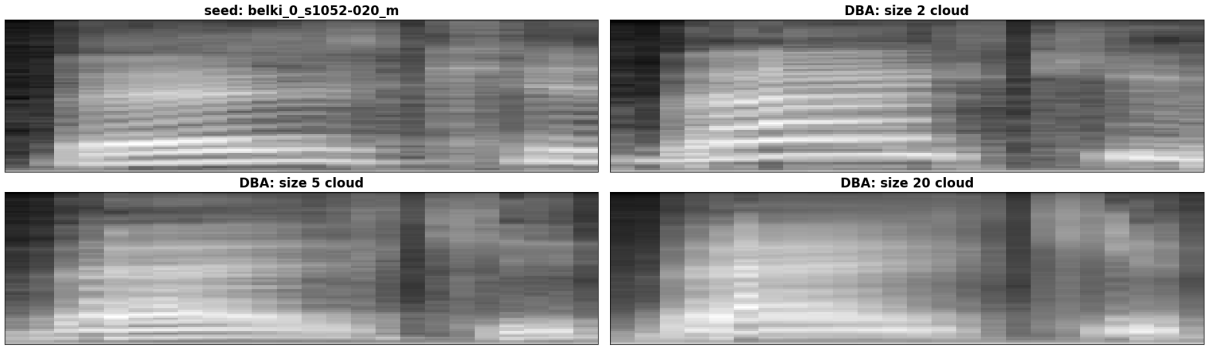

Figure 3: Output spectrogram of seed token of *belki* (“maybe”), along with spectrograms generated from 2, 5, and 20 tokens. Cloud size correlates with noisy outputs.

from 120 speakers (balanced across binarized gender categories; age 19–50 years, mean=23.9) who each read 40 sentences sampled from a triphone-balanced set of 2462 Turkish sentences (Özgül Salor et al., 2006). Metadata for each speaker includes (binarized) gender, dates of birth and recording, places of birth and residence, and level of education. Inspection revealed a subset (n=23) of the speakers in the corpus to have mismatches between audio and transcript files. These were filtered out, leaving 97 speakers (m=49, f=48) for all experiments described below.

Each recorded sentence is transcribed in standard Turkish orthography as well as an ASCII-compatible phonemic orthography derived from SAMPA (Wells, 1997), called *METUbet* (Özgül Salor et al., 2002). The corpus also includes word-level, phone-level, and HMM state alignments, computed with an HMM-GMM acoustic model trained on a subset of the full set of sentences.

As with most linguistic corpora, word frequencies follow a roughly Zipfian distribution. There are 7412 words in our dataset, the most frequent of which, *bir* (“one/a”), occurs approximately 897 times, whereas there are 2423 words which occur only once.

4.1 Model inputs

As mentioned, Mnemorphon’s inputs are words; these are segmented from the corpus speech files using the provided word-level alignments. Each segmented word is stored with its METUbet string representation as category label, along with speaker ID, gender marker, and a within-speaker token index. The segmented word audios are then encoded as mel-scaled spectrograms, with the following parameters:

- window length: 46ms
- hop length: 12ms
- 80 mel bands

As illustrated in Figure 2, these spectrogram parameters generate comparatively coarse *narrow-band* spectrograms, unlike e.g. Praat’s default values which have finer temporal resolution and are perhaps better suited to visual presentation. Our choice of spectrogram parameters was constrained by our evaluation methodologies, discussed below.

5 Experiment 1: cloud composition

In our initial experiments we explore the effect of cloud composition on Mnemorphon’s outputs. We begin with a maximally unconstrained approach, conditioning cloud selection solely on word category membership. For each of the word categories (i.e. distinct METUbet strings) represented in our corpus, we uniform randomly select one token as the seed exemplar and sample progressively larger uniform random subsets of the remaining tokens from the category as the cloud from which Mnemorphon computes a barycenter. We illustrate the outcome here in Figures 3 and 4 for a representative example, the form *belki* (gloss: “maybe”, corpus freq: 40, rank: 43). We plot full spectrograms and a selection of mel bands, respectively, for the seed token along with from Mnemorphon’s output barycenter for varying cloud sizes.

Figure 3 shows clearly that increasing the number of tokens included in the cloud results in Mnemorphon’s output spectrograms becoming “blurrier”, losing most of the fine structure present in individual tokens, particularly with respect to frequency information. Notwithstanding this noise, the individual mel bands plotted in Figure 4 show that Mnemorphon’s generation algorithm

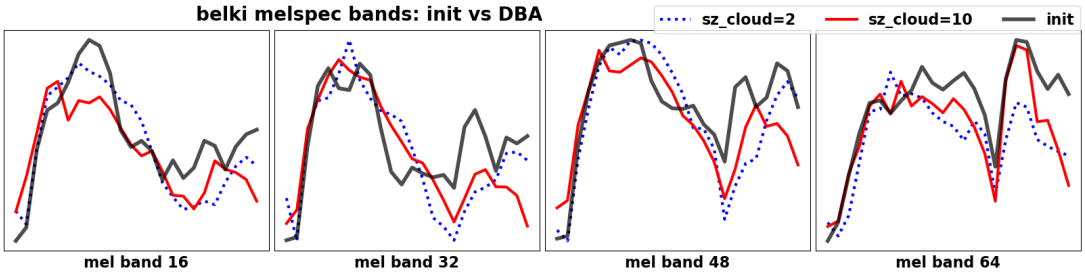

Figure 4: Mel bands 16, 32, 48, 64 of seed spectrogram, with DBA spectrograms from cloud size 2, and 10, for seed token *belki*.

does find meaningful averages for temporally variable signals, locating and aligning the major peaks and troughs in the energy for each band along the temporal dimension.

The relation between output noise and increasing numbers of cloud exemplars found here is not solely due to cloud size, but rather that the clouds' tokens are dispersed in the parametric space. To confirm this we generate outputs from the same seed, this time with two small clouds of the same size (N=3), constrained to contain the maximally similar and dissimilar tokens in the category, respectively. The outputs, shown in Figure 5, confirm that dispersion plays a key role in the quality of MNEMORPHON's generated forms. This in turn raises the question of latent categorical structure or organization within exemplars clouds.[8]

As discussed in Section 2, some of the early motivation for exploration of exemplar-based speech processing was the apparent storage and use of *non-linguistic* information, for example indexical information. The results above suggest that constraining MNEMORPHON's cloud selection by using any such additional contextually available information would likely serve to further reduce the output variance, resulting in cleaner, and in a sense more representative, output spectrograms. To test this we re-ran the same experiment as above, with the same seed token, this time constraining MNEMORPHON's clouds by using the (binarized) gender information that is available in our corpus. We used the output from our initial, unconstrained, experiment for a cloud size $N = 10$, and then used ten uniform randomly selected tokens from the relevant category that were tagged F ("female") in our corpus. Once again, we see in Figure 6 that constraining MNEMORPHON's cloud along dimensions of similarity, linguistic or otherwise, yields cleaner, more representative outputs.

Notwithstanding the obscuring or blurring of phonetic detail in MNEMORPHON's outputs, larger scale patterns of energy distribution across different frequency bands and time slices remain visible, hinting at an emergent, transient form of abstraction; a hallmark of exemplar models. In our next experiment we see that there is indeed linguistic categorical information recoverable from these outputs.

In addition to the direct visual evaluation here, we use a publicly available pre-trained neural vocoder (Lee et al., 2023)[9] to re-synthesize audio from our generated spectrograms for impressionistic auditory evaluation.[10] It is the use of this vocoder that constrained the spectrogram parameters in our data preparation; because BigVGAN is trained on narrow-band spectrograms (the standard choice in neural text-to-speech synthesis), these are required for any subsequent synthesis. That said, the finer frequency resolution of narrow-band spectrograms is likely beneficial for the quantitative evaluation in Experiment 6.

6 Experiment 2: latent categorical information in MNEMORPHON's outputs

We have seen that MNEMORPHON's outputs quickly become noisy as a function of cloud size, although this is somewhat mitigated by heavily weighting the influence of cloud tokens that are close to the seed in DTW distance. Despite this noise, we wish to determine whether generated outputs retain any categorically characteristic phonetic signal. We investigate this in the present experiment, in which we train a neural network to take spectral slices as inputs and classify them as *front* or *back* vowels.

[8] We thank an anonymous reviewer for highlighting this point and encouraging us to explore it.

[9] https://github.com/NVIDIA/BigVGAN

[10] The accompanying website hosts samples of audio synthesized from MNEMORPHON's outputs.

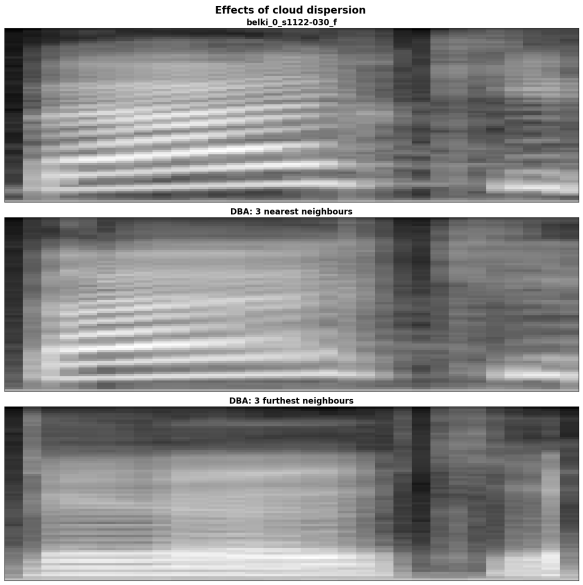

Figure 5: Effect of cloud dispersal; spectrogram of seed token of *belki*, along with spectrograms generated from clouds with minimal and maximal dispersion (nearest and furthest 3 neighbours, respectively).

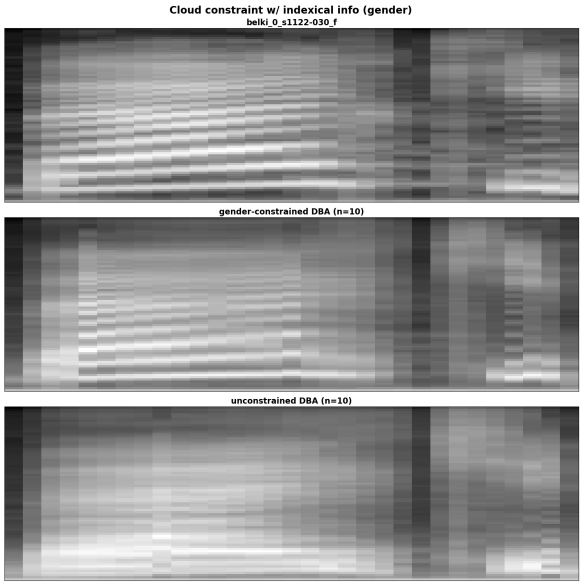

Figure 6: Effect of "gender"-based cloud constraint; spectrogram of seed token of *belki*, along with spectrograms generated from size $N = 10$ clouds restricted to tokens tagged as "female" versus sampled uniformly across gender markers.

Our focus on this particular phonetic characteristic, foreshadows work in progress assessing MNEMORPHON's ability to learn productive morphophonological generalizations, in particular Turkish front/back vowel harmony.[11]

Although MNEMORPHON itself has no notion of sub-lexical units, they are useful in the context of this extrinsic analysis. For this experiment, we extracted all vowels from the audio corpus using the included alignments, and converted them directly to mel spectrograms, resulting in a total of 82360 samples, which were randomly shuffled and divided via stratified split into train, development, and test sets representing 80, 10, and 10 percent of the corpus samples.

Our classifier is a convolution neural network. They are known to perform well on spectrograms and in fact form the backbone of many current speech recognition systems (Gulati et al., 2020). Our network has 4 layers of 2-d convolutions (5x5 in the first layer and 3x3 for subsequent layer), a max-pooling layer, and a final fully-connected layer projecting to a binary output (modeling [± back]). Kernel sizes, learning rate and batch size were tuned on a development split; the final training run was for 25 epochs.[12]

[11]Turkish also has rounding harmony, which we also leave for future investigation.

[12]See the accompanying repository for fuller details of the data generating process, network architecture, and training procedure.

6.1 Data augmentation

Like all supervised learning approaches, neural networks are sensitive to distribution shift, where the properties that the network learns to extract as relevant features are differently distributed in the training and evaluation sets. This exact situation obtains in the current experiment, where our training data consists solely of "clean" spectrograms directly computed from audio while the target spectrograms are "noisy" for reasons discussed above. For this reason our initial attempts at classifying MNEMORPHON's outputs fared poorly.

In order to mitigate the effect of this disparity we augmented our training data with DBA-generated samples; for each vowel category we added 1000 samples, each created by running distance-weighted DBA over 10 tokens uniform randomly sampled from the given category's exemplars in the training set.

6.2 Results

Table 1 shows the precision, recall, and F_1 score of our classifier on the test split of our data set. We can see that MNEMORPHON is, at least according to our classifier, producing output vowels with phonetic characteristics that enable their identification as front or back.

class	precision	recall	F_1	support
front	0.880	0.878	0.879	3154
back	0.866	0.867	0.867	2856
accuracy			0.873	6010

Table 1: Precision, recall, F_1 score, and accuracy of CNN phone classifier on held-out set

7 Discussion

We have shown here that *dynamic time warping* and *DBA barycenter averaging* together constitute a viable basis for a production algorithm in an exemplar model, MNEMORPHON, whose token representations are word-sized mel spectrograms of variable durations, overcoming a core challenge for exemplar production models. We showed both qualitatively and quantitatively that despite noise introduced by averaging over tokens that are dispersed in spectrotemporal space, our model's outputs retain phonetic properties that are characteristic of the exemplars from the generating categories.

8 Limitations and future work

MNEMORPHON's production algorithm as applied in these experiments generates comparatively noisy outputs, unless the selection of tokens for the exemplar cloud is severely constrained. Nonetheless, we see this work as an initial step toward a fully articulated theory and model of exemplar-based (psycho)linguistic knowledge. An eventual goal is to assess how far such a "pure" or "core" model can take us before a hybrid approach becomes necessary (cf. Goldrick and Cole, 2023).

In future work will explore further restrictions on cloud construction, exploring e.g. speaker identity, dialect, and speech rate among others.

As hinted in Section 6, we also intend to extend this work to account for productive morphophonological alternations like Turkish vowel harmony (see Mailhot, 2010b, for an exemplar production approach that learns productive vowel harmony on toy data, including patterns of opaque and transparent neutrality), and eventually to data from psycholinguistic research on speech perception and production (e.g. contexts of phonetic reduction and lengthening, and patterns of interlocutor convergence).

As the data used here are not widely accessible, we also intend to reproduce these results in the not-too-distant future using data from the Mozilla Common Voice corpus (Ardila et al., 2020)[13] in order to facilitate reproducibility.

8.1 A note on gender

As a final remark, we acknowledge here that gender identity and expression exist on a spectrum, and hence that the use of binarized gender in the experiment on constraining cloud size is problematic. The experiment was added in response to a pertinent reviewer remark, and in the interest of expediency we used the binarized gender markers that are available in our corpus's metadata. In future work we hope to address this more carefully, either using a wider array of self-reported gender identities, or potentially relying purely on phonetic features, e.g. high or low F_0 (although of course this is at best an approximation).

Acknowledgements

We would like to thank our reviewers for valuable feedback. Although this work has not been directly shared with any audiences, many of the ideas here are the result of hallway discussions and email exchanges with friends and colleagues over several years.

References

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

Jean Berko Gleason. 1958. The child's learning of english morphology. *Word*, 14.

Joan Bybee. 2001. *Phonology and Language Use*. Cambridge Studies in Linguistics. Cambridge University Press.

Li Deng and Douglas O'Shaughnessy. 2003. *Speech Processing: A dynamic and optimization-oriented approach*. Marcel Dekker, New York, NY.

Isaac Elias. 2006. Settling the intractability of multiple alignment. *Journal of computational biology : a journal of computational molecular cell biology*, 13(7):1323–1339.

Marc Ettlinger. 2007. An exemplar-based model of chain shifts. In *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS XVI)*.

[13] https://commonvoice.mozilla.org

Marc Ettlinger and Keith Johnson. 2010. Vowel discrimination by english, french and turkish speakers: Evidence for an exemplar-based approach to speech perception. *Phonetica*, 66(4):222–242.

Susanne Gahl and Alan C. L. Yu. 2006. Introduction to the special issue on exemplar-based models in linguistics. *The Linguistic Review*, 23(3):213–216.

Stephen Goldinger. 1996. Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning Memory and Cognition*, 22(5):1166–1183.

Stephen Goldinger. 1998. Echoes of echoes? an episodic theory of lexical access. *Psychological review*, 105:251–79.

Matthew Goldrick and Jennifer Cole. 2023. Advancement of phonetics in the 21st century: Exemplar models of speech production. *Journal of Phonetics*, 99.

Joshua T. Goodman. 2001. A bit of progress in language modeling. *Computer Speech & Language*, 15(4):403–434.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Proc. Interspeech 2020*, pages 5036–5040.

Sarah Hawkins. 2003. Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, 31(3):373–405. Temporal Integration in the Perception of Speech.

Douglas Hintzman. 1986. "schema abstraction"' in a multiple-trace memory model. *Psychological Review*, 93:411–428.

Keith Johnson. 1997a. The auditory/perceptual basis for speech segmentation. In *OSU Working Papers in Linguistics*, 50, pages 101–113. Ohio State University. Department of Linguistics.

Keith Johnson. 1997b. Speech perception without speaker normalization: an exemplar model. In K. Johnson and J.W. Mullenix, editors, *Talker Variability in Speech Processing*. Academic Press, San Diego.

Keith Johnson. 2006. Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics*, 34(4):485–499. Modelling Sociophonetic Variation.

Keith Johnson. 2007. Decisions and Mechanisms in Exemplar-based Phonology. In Maria-Josep Solé, Patrice Speeter Beddor, and Manjari Ohala, editors, *Experimental Approaches to Phonology*. Oxford University Press.

Robert Kirchner, R. Moore, and T-Y Chen. 2010. Computing phonological generalization over real speech exemplars. *Journal of Phonetics*, 38.

Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. 2023. BigVGAN: A universal neural vocoder with large-scale training. In *The Eleventh International Conference on Learning Representations*.

James MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley symposium on mathematical statistics and probability*, pages 281–297.

Frédéric Mailhot. 2010a. Instance-based acquisition of vowel harmony. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pages 1–8, Uppsala, Sweden. Association for Computational Linguistics.

Frédéric Mailhot. 2010b. *Modelling the acquisition and evolution of vowel harmony*. Ph.D. thesis, Carleton University.

Douglas L. Medin and Marguerite M. Schaffer. 1978. Context theory of classification learning. *Psychological Review*, 85(3):207–238.

Abdullah Mueen and Eamonn Keogh. 2016. Extracting optimal performance from dynamic time warping. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 2129–2130, New York, NY, USA. Association for Computing Machinery.

Robert M. Nosofsky. 1986. Attention, similarity, and the context theory of classification. *Journal of Experimental Psychology*, 115:39–57.

François Petitjean, A. Ketterlin, and P. Gançarski. 2011. A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, 44(3).

Janet B. Pierrehumbert. 2001. Exemplar dynamics: Word frequency, lenition and contrast. *Typological Studies in Language*, 45:1–11.

Janet B. Pierrehumbert. 2002. Word-specific phonetics. In *Laboratory Phonology 7*, pages 101–140, Berlin, New York. De Gruyter Mouton.

H. Sakoe and S. Chiba. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49.

Richard Semon. 1923. *Mnemic Psychology*. George Allen & Unwin, London. Translated by B. Duffy (original work publish 1909).

Romain Tavenard, Johann Faouzi, Gilles Vandewiele, Felix Divo, Guillaume Androz, Chester Holtz, Marie Payne, Roman Yurchak, Marc Rußwurm, Kushal Kolar, and Eli Woods. 2020. Tslearn, a machine learning toolkit for time series data. *Journal of Machine Learning Research*, 21(118):1–6.

Taras K. Vintsyuk. 1968. Speech discrimination by dynamic programming. *Cybernetics*, 4:52–57.

Andrew Wedel. 2006. Exemplar models, evolution and language change. *The Linguistic Review*, 23(3):247–274.

John Wells. 1997. Sampa - computer readable phonetic alphabet. Accessed on January 20, 2024.

Özgül Salor, Bryan Pellom, Tolga Çiloglu, Kadri Hacioglu, and Mübeccel Demirekler. 2002. On developing new text and audio corpora and speech recognition tools for the turkish language. In *Proc. 7th International Conference on Spoken Language Processing (ICSLP 2002)*, pages 349–352.

Özgül Salor, Bryan Pellom, Tolga Çiloglu, Kadri Hacioglu, and Mübeccel Demirekler. 2006. Middle east technical university turkish microphone speech (v1.0) ldc2006s33. Web download. Philadelphia; Linguistic Data Consortium.

A PEBLS : Phonological Exemplar-based Learning System

Kirchner et al. (2010) present PEBLS, to our knowledge the only exemplar production model in the phonetics/phonology literature that operates over (digitized representations of) real speech tokens.

To produce an output, PEBLS randomly selects a seed token from the set of word labels; all remaining exemplars in that set serve as the cloud. Output production is then cast as the problem of determining an optimal alignment between the seed and the entire cloud. Concretely, PEBLS's output is a token composed of coordinates or sub-sequences of in-cloud exemplars that may occur in *any position* in *any token*. The optimization is computed over both coordinate-wise similarities, and inter-coordinate transition similarities (these obtained by computing an alignment of the cloud with itself, offset by one coordinate.)

Kirchner et al. note that this production method also faces the issue of generalization, as for categories whose exemplars mostly-with-exceptions reflect some phonological generalization (e.g. intervocalic lenition). If the initially sampled seed token violates the relevant generalization (i.e., it includes a stop between vowels), and even a single generalization-violating exemplar exists in the cloud, it will be directly output by PEBLS, notwithstanding the preponderance of generalization-conforming exemplars.

In order to predispose PEBLS to produce tokens that reflect the statistical generalizations instantiated in its exemplars, a "confidence" measure is introduced that expresses the representativeness of sequences of coordinate transitions within the cloud. This confidence computation requires a complete hierarchical clustering over the cumulative partial DTW scores at each coordinate transition.

While PEBLS presents solutions to the problems of production and generalization over real speech exemplars, it does so at the cost of non-trivial complexity; introducing unmotivated modifications to the DTW algorithm, along with an ad-hoc mechanism to down-weight the importance of non-representative exemplars within a cloud.

Japanese Rule-based Grapheme-to-phoneme Conversion System and Multilingual Named Entity Dataset with International Phonetic Alphabet

Yuhi Matogawa[r], Yusuke Sakai[r], Taro Watanabe[r], Chihiro Taguchi[I]
[r]Nara Institute of Science and Technology, [I]University of Notre Dame
{matogawa.yuhi.na2, sakai.yusuke.sr9, taro}@is.naist.jp, ctaguchi@nd.edu

Abstract

In Japanese, loanwords are primarily written in Katakana, a syllabic writing system, based on their pronunciation. However, the transliterated loanwords often exhibit spelling variations, such as the word "Hepburn" being written as "ヘボン (hebon)", "ヘプバーン (hepubaan)", "ヘップバーン (heppubaan)". These orthographical variants pose a bottleneck in multilingual Named Entity Recognition (NER), because named entities (NEs) do not have one-to-one matches. In this study, we introduce a rule-based grapheme-to-phoneme (G2P) system for Japanese based on literature in linguistics and a large-scale multilingual NE dataset with annotations of the International Phonetic Alphabet (IPA), focusing on IPA to address the Katakana spelling variations in loanwords. These rules and dataset are expected to be beneficial for tasks such as NE aggregation, G2P system, construction of cross-lingual language models, and entity linking. We hope our work advances research on Japanese NER with multilingual loanwords by solving the spelling ambiguities[1].

1 Introduction

Japanese orthography consists of three unique writing systems: Hiragana, Katakana, and Kanji. Among these, Katakana is mainly used for transliteration of loanwords originating from languages outside the Sinosphere, in particular, for named entities (NEs) such as proper nouns. However, there are no clear rules for this transliteration process, and NEs are transliterated into the closest Katakana based on the pronunciation of the source language. While some loanwords close to the Japanese pronunciation are often transliterated into Katakana representing mostly similar sounds (e.g., "Obama" to "オバマ (obama)"), this is not the case for other loanwords, leading to ambiguities due to the lack of unified common transliteration (e.g., "Hepburn" to "ヘボン (hebon)", "ヘプバーン (hepubaan)", and "ヘップバーン (heppubaan)"). More details are described in Appendix A.

This ambiguity poses challenges in unifying Katakana-written loanword NEs within Japanese and identifying the original NEs. This issue stems from significant differences in phonological and writing systems between Japanese and other languages, especially in historical documents written when the reading of foreign words was not customary among general readers, resulting in a wide variety of transliterations based on the sound perception of individuals without established transliteration rules. Furthermore, the inconsistent transliteration of loanword NEs in Japanese can be an obstacle for Japanese learners.

Given these issues, we first developed a rule-based Japanese Grapheme-to-Phoneme (G2P) system. Although Katakana can be mapped into IPA by rule-based conversion, the development is challenging because constructing precise rules requires knowledge of linguistics and phonetics. Currently, only neural-based approaches support Japanese G2P. Our rule-based G2P system, founded on linguistic principles, accurately represents pronunciation even in cases where IPA is automatically extracted from a source whose phonetic accuracy is not guaranteed. We hope that our G2P system can serve as a useful learning aid for Japanese learners and is expected to be applicable to pronunciation-based Japanese text analysis methods, such as NER and entity linking.

Next, we constructed a large-scale multilingual NE dataset with IPA annotations to address the spelling variations such as Katakana spellings of loanwords in Japanese. This dataset contains over 69 million pairs of NE and IPA, and over 14 million

[1]The original code and dataset are available from https://github.com/lart-rt/Japanese_ipa_rule_and_NE_dataset. Moreover, our work is merged with Epitran (Mortensen et al., 2018), a widely-used G2P library: https://github.com/dmort27/epitran/pull/143. You can access the demo site on https://yusuke1997.com/Japanese_G2P.

Proceedings of the The 21st SIGMORPHON workshop on Computational Morphology, Phonology, and Phonetics, pages 77–86

IDs used to identify NEs from 68 languages. On average, each ID is associated with five different pairs of NE and IPA. Ours is the largest dataset among the multilingual NE datasets with phonetic annotation. We hope our work advances research dealing with the spelling variations in Japanese, including approaches of cross-lingual word alignments such as Knight and Graehl (1997); Ren (2023), which utilizes phoneme sequences (though non-IPA) as an intermediate to align English and Japanese words.

To summarize, our contributions are as follows:

- We developed a rule-based G2P conversion system from the Japanese linguistic literature.
- We constructed a large-scale multilingual NE dataset with IPA annotations considering the transliteration ambiguity.

2 Related Work

G2P Conversion Systems. G2P conversion systems are mainly classified into two types: rule-based (Pine et al., 2022; Sar and Tan, 2019; Kłosowski, 2022; Deri and Knight, 2016; Wang and Tsai, 2009; Alam et al., 2011; Narasimhan et al., 2004) and neural-based (Li et al., 2022; Yamasaki, 2022; Peters et al., 2017; Arora et al., 2020) / machine learning-based (Rama et al., 2009; Laurent et al., 2009; Kienappel and Kneser, 2001) approaches. The neural-based approaches achieve high performance for high-resource languages, but the performance is significantly degraded when the quality of phonemic representation in the training data is not ensured or when the dataset size is small (Clematide and Makarov, 2021). On the other hand, rule-based approaches can easily obtain accurate G2P results and are faster than neural-based approaches, when rules are built on correct pronunciation based on linguistic features.

Japanese G2P. There have been attempts to develop Japanese systems for G2P or grapheme-phoneme alignment using neural-based (Makarov and Clematide, 2020; Clematide and Makarov, 2021; ElSaadany and Suter, 2020; Vesik et al., 2020) and machine learning-based approaches (Waxmonsky and Reddy, 2012; Bhargava and Kondrak, 2011; Baldwin and Tanaka, 1999; Nagata, 2000). The systems are applied to tasks such as estimating pronunciation in Japanese (Hatori and Suzuki, 2011; Yencken and Baldwin, 2005) and transliterating named entities (Tsuji et al., 2012; Bilac and Tanaka, 2004; Yamashita et al., 2018; Ren, 2023). However, these systems require training data, which are often not from sources of ensured quality, bearing the possibility of predicting incorrect IPA[2]. They are not based on linguistic insights, and although they have achieved success in some tasks of natural language processing, such a problem remains that they do not reflect truly correct pronunciation from the perspective of linguistics and phonetics. For rule-based approaches (Bilac et al., 1999; Shiga and Kawai, 2012; Terada and Lee, 2017; Masuda and Umemura, 1997; Sagisaka and Sato, 1983), the rules for G2P or grapheme-phoneme alignment that reflect accurate pronunciation based on literature about Japanese phonetics and phonology have not been published in any academic paper or presentation yet. The difficulty of Japanese G2P can be attributed not only to the complexity of its writing systems, which employ the three different systems (i.e., Hiragana, Katakana, and Kanji), but also to the syllabic characteristics of Hiragana and Katakana. More details are described in Appendix B.

Datasets for IPA. There are several multilingual datasets that match NE sequences in their original language with their corresponding IPA representations. However, they are not necessarily suitable for downstream tasks such as NER assumed in this study. WikiPron, for example, includes IPA from non-NE entries and is not readily applicable for solving tasks related to NEs. Klumpp et al. (2022) adds annotation of IPA to speech in six languages, but it still does not suit our purpose because it is also not specific to NE.

3 Building the Japanese Rule-based G2P System

3.1 Creating the rules

We created the G2P rules that reflect the description of Japanese phonetics and phonology (NKG, 2005; Saito, 2006). Specifically, our system was developed based on the chapter on phonetics and phonology in (NKG, 2005), the encyclopedia of

[2]For example, WikiPron (Lee et al., 2020): https://github.com/CUNY-CL/wikipron, the data regarded as gold in Ashby et al. (2021), is the corpus which comprises pairs of graphemes and IPA automatically extracted from online dictionary "Wiktionary": https://www.wiktionary.org. However, the correctness and consistency of these IPA representations are not guaranteed because these IPA are manually annotated by Wiktionary users, including non-experts of unknown academic backgrounds.

Japanese language education. In addition, we referred to the phonetic description in (Saito, 2006) to take into account some peculiar phonetic realizations. Additionally, we also created a simpler version of G2P rules based on other references. Upon implementing the G2P rules, the format conforms to the notation of the existing multilingual G2P framework Epitran [3] (Mortensen et al., 2018). Our work is the first contribution to G2P conversion for a language with syllabary scripts in the framework of Epitran.

The mechanism of Epitran Epitran has three types of conversion rules: "map", which exists for all languages, and "pre" and "post", which are optional to some languages. The basic conversion is done by "map", which is a one-to-one correspondence between the letters of each language and the IPA symbols, but "pre" and/or "post" are applied before and/or after "map", respectively, as needed to handle phenomena that cannot be handled by the one-to-one correspondence, such as when the pronunciation changes depending on the environment of the preceding and/or following sounds. If necessary, "pre" and/or "post" are applied before and/or after "map" is applied, respectively. For example, German "ö" is basically converted to [ø] according to "map", though to [œ] when two or more consonants follow it due to "post".

Among the three types of rules in Epitran, we created "map" and "post" for Japanese because all the pronunciation mappings can be done by these two in the language. In other words, we created "map" for each combination of Katakana / Hiragana and an IPA symbol, and "post" to deal with phenomena that cannot be handled by "map". Table 1 shows the examples of both in the detailed version of the rules.

Detailed version of rules In our work, the criteria for IPA granularity prioritize phonetic accuracy over phonemic representation, without exceeding what is necessary. While phonetically accurate transcription is important, perfectly phonetic representation can be not only redundant but also impractical, since we do not have spoken data. For our purpose, the description in NKG (2005) meets these criteria. It was originally published for Japanese language education and includes linguistically precise descriptions of various aspects of the Japanese language including phonetics and phonology, which we mainly referred to in this study.

[3] https://github.com/dmort27/epitran

Type	Conversion rules	Conditions
map	ラ ⇒ ɾa	–
post	ɾ ⇒ ɖ / # _	if word-initial

Table 1: Examples of "map" and "post" in Japanese G2P rule.

We created "map" mostly based on basic mappings between Katakana and IPA described in NKG (2005). However, the pronunciation of characters can vary depending on their surrounding environment. For such cases, we incorporated such phonetic variations in "post", drawing from the description of phonetics and phonology of Japanese given in the literature. To cover the phonetic rules comprehensively, we also referred to the other work Saito (2006), which provide a more detailed description of Japanese phonetics than NKG (2005). For instance, the moraic nasal /N/ (Katakana: "ン") has different phonetic realizations [m], [n], or [ŋ] depending on the articulation of the following consonant. We describe this phenomenon using "post". Namely, we write the rules to update [ɴ] (the tentative IPA symbol for /N/ in "map") to either [m], [n], or [ŋ] conditioned by its succeeding sound.

After creating "map" and "post" for Katakana, we created the rules for Hiragana by converting Katakana to Hiragana. This is simply because Katakana and Hiragana have a complete one-to-one correspondence with each other.

Simplified rules In addition to the fine-grained rules for the mapping from Katakana to IPA, we prepared a set of simplified G2P rules based on the articles related to Japanese writing systems and phonology in the English Wikipedia.[4] Specifically, the number of the more accurate rules of "map" is 150 while for the simplified rules 112. Also for "post", the more accurate rules comprise 46 whereas the more simplified rules are 20.

The simplified rules are more phonemic and phonetically less fine-grained than the detailed rules. However, creating the simplified version allows users to have multiple choices; for example, a user

[4] "Katakana - Wikipedia" (https://en.wikipedia.org/w/index.php?title=Katakana&oldid=1103341275, viewed in 2022 August) and "Sokuon - Wikipedia" (https://en.wikipedia.org/w/index.php?title=Sokuon&oldid=1096454475, viewed in 2022 August. "sokuon" means the first part of a geminated consonant).

may only want a reduced system for IPA without phonetic details. The simplified version also includes the mapping rules for both Katakana and Hiragana.

3.2 Evaluation

We evaluate the G2P conversion with our rules in comparison to the one by WikiPron. WikiPron is constructed by automatic crawling from Wiktionary and thus the quality of IPA conversion is not ensured. Some of them seem linguistically incorrect and different from the actual pronunciation. On the other hand, our rules can reflect correct pronunciation even in these cases because our rules are fully based on the literature in linguistics and phonetics. We compare IPA in WikiPron to IPA converted by our rules and show our rules are more preferable when IPA in WikiPron is wrong.

We used 2,348 Katakana–IPA pairs in WikiPron and compared WikiPron's IPA to IPA converted by our rules from Katakana in the dataset. We show the patterns of differences between WikiPron and ours in Table 2. As shown in (a) and (b) in the table, WikiPron incorrectly represents pronunciation for more than a quarter of the words. For instance, word-initial /r/ is transcribed as [ɾ] in WikiPron, though it is inappropriate according to (Saito, 2006). In contrast, ours converts word-initial /r/ to [ɖ], as pointed out in the literature, representing the appropriate pronunciation. We cannot judge which is more correct between WikiPron and ours in pattern (c), while in (d) ours are wrong. However, note that the proportion of (a) and (b), the pattern where WikiPron's is wrong, is much higher than pattern (d) where ours are inappropriate. Moreover, the cases in (d) are only limited to either of the following, both of which are highly exceptional in Japanese:

- When the word includes a less frequent or seldom used mora. These moras are not native to Japanese phonology but originate from foreign languages (e.g. "ヴャ (vya)").
- When the word is written in the archaic orthography that is no longer used in modern Japanese. In addition to the example in Table 2, an interesting instance is "シャッター". The literal pronunciation is "shiyattaa", though it is actually pronounced as "shattaa". The word is now almost totally replaced with "シャッター", which Japanese speakers also read as "shattaa".

		Examples		
Pattern	# notations	Katakana	WikiPron	Ours
(a)	673 (28.7%)	ラム	[ramɯ]	[ɖamɯ]
(b)	5 (0.213%)	ユータナジー	[oitanaʑiː]	[juːtanaʑiː]
(c-1)	1679 (71.5%)	ディスコ	[dʲisɯko]	[disɯko]
(c-2)	528 (22.5%)	ベンチ	[bẽɲtɕi]	[bentɕi]
(d)	36 (1.53%)	ヰ	[i]	"ヰ"

Table 2: The comparison of WikiPron and our rule-based system. (a) means some sounds represented in WikiPron are wrong according to the literature. (b) indicates IPA in WikiPron actually represents different Katakana from what is aligned with the IPA in the dataset, where the Katakana is a variant of the same word (the example in this table belongs to this type) or the Katakana is a completely different word. Pattern (c) refers to the cases where it is not clear whether the sound in WikiPron is wrong based on the description given in the literature. Among (c), WikiPron and ours differ in supplemental symbols in (c-1) and in main symbols in (c-2). Pattern (d) includes cases where some characters in Katakana are not supported in our rules. Note that the sum is not 100% since one sample can include multiple error patterns. For instance, IPA of "リンチ (rinchi)" falls into both of patterns (a) and (c).

4 The multilingual dataset of pairs of NE and IPA

4.1 Constructing the dataset

In addressing the challenge of Japanese NE variants, we also constructed a multilingual NE dataset. This dataset comprises pairs of NEs and their respective IPA representations, derived from the NE dataset ParaNames[5](Sälevä and Lignos, 2022). We achieved this by converting NEs of each language using Epitran for each language and Japanese NEs using the Japanese rules we introduced in Section 3. There are also other multilingual datasets of NEs such as "TRANSLIT" (Benites et al., 2020), but we chose ParaNames because it has the largest size and the widest coverage of languages. We created 69 million pairs in 68 languages by leveraging G2P rules in Epitran for each language, in which approximately 671K pairs were Japanese.

ParaNames entirely derives from the structured knowledge base of entries in Wikipedia. Specifically, NEs registered in Wikidata as instances of either "human", "geographic region", or "organization" are extracted for each language supported in Wikipedia. One set of data consists of "wikidata_id"

[5]https://github.com/bltlab/paranames

	Key	Value		
ParaNames	wikidata_id	Q19618413		
	type	LOCATION		
	language	English	Chinese	Japanese
	label	Paris	巴黎	パリ
+ Ours	*ipa*	**pɛʁis**	**pali**	**paɾji**

Table 3: An Example of ParaNames and our contributions

for the ID given in Wikidata, "label" for the notation of the NE, "language" for the language tag associated with the notation in "label", and "type" for the type of the NE. Appendix C describes more details about ParaNames entries.

We converted notations in "label" columns to IPA by Epitran for 68 languages and added IPA to the original data as shown in Table 3. The rules for most languages consist of some or all of "map", "pre", and "post" except English and Chinese, for which it is difficult to implement the one-to-one G2P mapping. For this reason, we leverage external pronunciation dictionaries for these two exceptional languages: "flite"[6] for English and "CC-CEDICT"[7] for Chinese.

4.2 Statistics of the dataset

Table 4 presents the overview of the statistics of our dataset. The pairs of NE–IPA amounts to more than 69 million and the number of IDs associated with pairs is over 14 million. This results in 4.964 pairs of NE-IPA per ID on average. The average number of pairs of notations per language tag and per language is 732K and 1.023 million, respectively. This size is much larger than the existing dataset with graphemes and phonemes of WikiPron, which has only about 14K pairs per language. This suggests the effectiveness of this dataset when applied to linking or alignment between different notations of the same NE. Therefore, the dataset we constructed in this study is distinguished from the multilingual datasets with IPA in the previous studies in that not only it is specific to NEs but also the amount of data per language is much greater than that of existing datasets.

The type of NE with the highest frequency is PER (person). The average sequence lengths of the original NE and IPA are 15.085 and 15.894, respectively. The number of notations per language tag are provided in Appendix D.

The number of writing systems of original NE is 20 in total[8]. 15 languages have more than one million pairs of NE-IPA, approximately over average per language, for each. All of them use Latin script except Russian and Chinese while all of these belong to Indo-European except Hungarian and Chinese.

Measurement	Value
Total number of notations	69,573,951
– PER	48,625,240
– LOC	13,905,603
– ORG	7,043,108
Total number of IDs	14,016,907
– PER	8,897,440
– LOC	3,464,982
– ORG	1,654,485
Number of language tags	95
Number of actual languages	68
Average character length of NE notations	15.085
Average character length of IPA sequences	15.894

Table 4: Key statistics of our dataset.

5 Conclusion

In this paper, we introduced the new G2P rules for Japanese based on the literature in linguistics and phonetics, and constructed the largest multilingual dataset of NE with IPA using rule-based G2P including our own rules for Japanese. These resources will be beneficial for solving NE-related tasks such as NER and entity linking in Japanese. The actual application of our G2P tool and dataset to downstream tasks on Japanese NEs like transliteration, text-to-speech, and so on is left for future work.

6 Limitations

G2P conversion system While our research primarily focused on the development of G2P for Katakana, we have also made it compatible with Hiragana and Kanji as a prototype, allowing for the input of any Japanese text. However, the complexity of Kanji readings far exceeds our initial estimations, making full support a future challenge.

[6] "flite: A small fast portable speech synthesis system" (https://github.com/festvox/flite, viewed in 2023 January).

[7] "CC-CEDICT Home [CC-CEDICT WIKI]" (https://cc-cedict.org/wiki/#what_is_cc-cedict, viewed in 2023 January).

[8] Latin, Ge'ez, Arabic, Cyrillic, Bengali, Devanagari, Katakana, Hiragana, Khmer, Rao, Malayalam, Burmese, Oriya, Gurmukhi, Sinhalese, Tamil, Telugu, Thai, simplified Chinese, and traditional Chinese.

Nonetheless, we have confirmed that G2P conversion is possible with a general level of accuracy.

Dataset In this paper, our contributions are the development of the NE dataset with IPA annotations and do not include experiments in downstream tasks. However, applying NE datasets with IPA annotations to downstream tasks has been reported in recent studies (Hentona et al., 2022), and we intend to apply ours to downstream tasks in future work. Additionally, the development of our large-scale dataset, comprising over 69 million NEs with associated IPA representations, demanded a significant investment of computational resources and time. We used a total of 96 CPU cores of Intel(R) Xeon(R) Platinum 8160 CPU @ 2.10GHz and 384GB RAM, taking nearly two months to complete annotating IPA. The availability of the large-scale dataset enables rapid experimentation, rendering it a highly valuable resource for advancing future research.

7 Impact

7.1 Effectiveness of our Japanese G2P conversion system

Our rule-based G2P conversion system is based on linguistic literature on Japanese, as mentioned in Section 3, allowing it to perform accurate G2P transliteration. Furthermore, a rule-based conversion system enables faster transliteration than neural-based systems. Furthermore, as shown in Figure 1, we have launched a demonstration site that supports both PC and mobile environments to allow anyone to easily use our system. This effort is groundbreaking because it is not supported by existing G2P systems such as Epitran. Our demonstration site also supports most of Japanese characters, including Kanji, Katakana, and Hiragana. We hope that it will be utilized as a tool for learners of Japanese to accurately predict pronunciations. We plan the demonstration site to be supported long-term, with plans including OCR and mobile-native support, as well as expansion to other languages in the future.

7.2 Effectiveness of our dataset

When comparing the other datasets containing IPA or writing systems such as Katakana or Latin script, our dataset has mainly two benefits.

Phonetic accuracy. IPA in this dataset takes into account more accurate pronunciation than the one in WikiPron and simple romanization, since the rules were based on the specialized literature in phonetics. For example, “ラザフォード (razafoodo)” is transcribed as [ɖazafoːdo] in our system unlike inappropriate [razafoːdo] in WikiPron-like manner.

Figure 1: The screenshot of our Japanese G2P demonstration site. Index 0 in the results is written using Katakana, index 1 is written using Hiragana, and index 2 is written using Kanji characters. We input some Japanese words in the text input area, then output each IPA as G2P results. We mainly support Katakana, but any Japanese characters are accepted. We support both PC and mobile environments and continue to improve and ensure long-term support, so we hope our G2P site helps non-native or Japanese learners to know the pronunciation of Japanese words.

Smaller edit distance. IPA can contribute to identifying different notations of Katakana for one ID as the same entity with less cost than using the original Katakana. Given one entity with two different forms “ラザフォード” and “ラザホード (razahoodo)”, the edit distance between them in IPA is only 1 ([ɖazafoːdo] for the former and [ɖazahoːdo] for the latter) while the edit distance for Katakana is 2.

This multilingual dataset includes not only NEs widely acknowledged in Japanese (e.g. “Paris”, “Madrid”, etc.) but also NEs hardly used in Japanese. Thus, our dataset can link or align NEs in Katakana even when they are rare words.

References

Firoj Alam, S.M. Murtoza Habib, and Mumit Khan. 2011. Bangla text to speech using festival. In *Proceedings of Conference on Human Language Technology for Development*, pages 154–161. Bibliotheca Alexandrina.

Aryaman Arora, Luke Gessler, and Nathan Schneider. 2020. Supervised grapheme-to-phoneme conversion of orthographic schwas in Hindi and Punjabi. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7791–7795, Online. Association for Computational Linguistics.

Lucas F.E. Ashby, Travis M. Bartley, Simon Clematide, Luca Del Signore, Cameron Gibson, Kyle Gorman, Yeonju Lee-Sikka, Peter Makarov, Aidan Malanoski, Sean Miller, Omar Ortiz, Reuben Raff, Arundhati Sengupta, Bora Seo, Yulia Spektor, and Winnie Yan. 2021. Results of the second SIGMORPHON shared task on multilingual grapheme-to-phoneme conversion. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 115–125, Online. Association for Computational Linguistics.

Timothy Baldwin and Hozumi Tanaka. 1999. The applications of unsupervised learning to Japanese grapheme-phoneme alignment. In *Unsupervised Learning in Natural Language Processing*.

Fernando Benites, Gilbert François Duivesteijn, Pius von Däniken, and Mark Cieliebak. 2020. TRANSLIT: A large-scale name transliteration resource. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3265–3271, Marseille, France. European Language Resources Association.

Aditya Bhargava and Grzegorz Kondrak. 2011. How do you pronounce your name? improving G2P with transliterations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 399–408, Portland, Oregon, USA. Association for Computational Linguistics.

Slaven Bilac, Timothy Baldwin, and Hozumi Tanaka. 1999. Incremental japanese grapheme-phoneme alignment (日本語における漸進型書記素・音素アラインメント). *The Special Interest Group Technical Reports of IPSJ. Technical Reports of NL, IPSJ Natural Language Processing*, 130:9–16.

Slaven Bilac and Hozumi Tanaka. 2004. A hybrid back-transliteration system for Japanese. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 597–603, Geneva, Switzerland. COLING.

Simon Clematide and Peter Makarov. 2021. CLUZH at SIGMORPHON 2021 shared task on multilingual grapheme-to-phoneme conversion: Variations on a baseline. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 148–153, Online. Association for Computational Linguistics.

Aliya Deri and Kevin Knight. 2016. Grapheme-to-phoneme models for (almost) any language. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 399–408, Berlin, Germany. Association for Computational Linguistics.

Omnia ElSaadany and Benjamin Suter. 2020. Grapheme-to-phoneme conversion with a multilingual transformer model. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 85–89, Online. Association for Computational Linguistics.

Jun Hatori and Hisami Suzuki. 2011. Japanese pronunciation prediction as phrasal statistical machine translation. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 120–128, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Asahi Hentona, Takamichi Toda, Yuta Tomomatsu, Masakazu Sugiyama, Yuki Azuma, and Sho Shimoyama. 2022. Unsupervised entity linking based on word alignment considering similarity about word embeddings phoneme sequences (単語の分散表現および音素列の類似性を考慮した単語アラインメントに基づく教師なしEntity Linking). In *Proceedings of the 28th Annual Conference of the Association for Natural Language Processing*, pages 1568–1572. Association for Natural Language Processing. In Japanese.

Anne K. Kienappel and Reinhard Kneser. 2001. Designing very compact decision trees for grapheme-to-phoneme transcription. In *Proceedings of 7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, pages 1911–1914. International Speech Communication Association.

Philipp Klumpp, Tomas Arias, Paula Andrea Pérez-Toro, Elmar Noeth, and Juan Orozco-Arroyave. 2022. Common phone: A multilingual dataset for robust acoustic modelling. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 763–768, Marseille, France. European Language Resources Association.

Kevin Knight and Jonathan Graehl. 1997. Machine transliteration. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 128–135, Madrid, Spain. Association for Computational Linguistics.

Piotr Kłosowski. 2022. A rule-based grapheme-to-phoneme conversion system. *Applied Sciences*, 12(5).

Antoine Laurent, Paul Deléglise, and Sylvain Meignier. 2009. Grapheme to phoneme conversion using an

smt system. In *Proceedings of Interspeech 2009*, pages 708–711. International Speech Communication Association.

Jackson L. Lee, Lucas F.E. Ashby, M. Elizabeth Garza, Yeonju Lee-Sikka, Sean Miller, Alan Wong, Arya D. McCarthy, and Kyle Gorman. 2020. Massively multilingual pronunciation modeling with WikiPron. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4223–4228, Marseille, France. European Language Resources Association.

Xinjian Li, Florian Metze, David Mortensen, Shinji Watanabe, and Alan Black. 2022. Zero-shot learning for grapheme to phoneme conversion with language ensemble. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2106–2115, Dublin, Ireland. Association for Computational Linguistics.

Peter Makarov and Simon Clematide. 2020. CLUZH at SIGMORPHON 2020 shared task on multilingual grapheme-to-phoneme conversion. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 171–176, Online. Association for Computational Linguistics.

Keiko Masuda and Kyoji Umemura. 1997. Extracting kana - alphabet rules from a non - japanese name reading table (人名辞書から名前読み付与規則を抽出する試み). *The Special Interest Group Technical Reports of IPSJ. Technical Reports of NL, IPSJ Natural Language Processing*, 69:97–102. In Japanese.

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Masaaki Nagata. 2000. Synchronous morphological analysis of grapheme and phoneme for Japanese OCR. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 384–391, Hong Kong. Association for Computational Linguistics.

Bhuvana Narasimhan, Richard Sproat, and George Kiraz. 2004. Schwa-deletion in hindi text-to-speech synthesis. *International Journal of Speech Technology*, 7:319–333.

Nihongo Kyoiku Gakkai NKG. 2005. *Encyclopedia of Japanese Language Education (new edition)*. Taishukan Shoten. In Japanese.

Ben Peters, Jon Dehdari, and Josef van Genabith. 2017. Massively multilingual neural grapheme-to-phoneme conversion. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 19–26, Copenhagen, Denmark. Association for Computational Linguistics.

Aidan Pine, Patrick William Littell, Eric Joanis, David Huggins-Daines, Christopher Cox, Fineen Davis, Eddie Antonio Santos, Shankhalika Srikanth, Delasie Torkornoo, and Sabrina Yu. 2022. G_i2P_i rule-based, index-preserving grapheme-to-phoneme transformations. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 52–60, Dublin, Ireland. Association for Computational Linguistics.

Taraka Rama, Anil Kumar Singh, and Sudheer Kolachina. 2009. Modeling letter-to-phoneme conversion as a phrase based statistical machine translation problem with Minimum Error Rate training. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Student Research Workshop and Doctoral Consortium*, pages 90–95, Boulder, Colorado. Association for Computational Linguistics.

Yuying Ren. 2023. Back-transliteration of English loanwords in Japanese. In *Proceedings of the Workshop on Computation and Written Language (CAWL 2023)*, pages 43–49, Toronto, Canada. Association for Computational Linguistics.

Yoshinori Sagisaka and Hirokazu Sato. 1983. Accentuation rules for japanese word concatenation (日本語単語連鎖のアクセント規則). *The IEICE Transactions*, J66-D:849–856. In Japanese.

Yoshio Saito. 2006. *Introduction to Japanese Phonetics (revised)*. Sanseido. In Japanese.

Jonne Sälevä and Constantine Lignos. 2022. ParaNames: A massively multilingual entity name corpus. In *Proceedings of the 4th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 103–105, Seattle, Washington. Association for Computational Linguistics.

Vathnak Sar and Tien-Ping Tan. 2019. Applying linguistic g2p knowledge on a statistical grapheme-to-phoneme conversion in khmer. *Procedia Computer Science*, 161:415–423. The Fifth Information Systems International Conference, 23-24 July 2019, Surabaya, Indonesia.

Yoshinori Shiga and Hisashi Kawai. 2012. Multilingual speech synthesis system. *Journal of the National Institute of Information and Communications Technology*, 59:21–28.

Takuya Terada and Akinobu Lee. 2017. Automatic construction of a robust pronunciation dictionary for spoken language using statistical learning by g2p in japanese (日本語におけるG2Pによる統計的学習を用いた話し言葉に頑健な発音辞書の自動構築). *The Special Interest Group Technical Reports of IPSJ. Technical Reports of SLP, IPSJ Spoken Language Processing*, 11:1–6. In Japanese.

Rieko Tsuji, Yoshinori Nemoto, Wimvipa Luangpiensamut, Yuji Abe, Takeshi Kimura, Kanako Komiya, Koji Fujimoto, and Yoshiyuki Kotani. 2012. The transliteration from alphabet queries to Japanese product names. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 456–462, Bali, Indonesia. Faculty of Computer Science, Universitas Indonesia.

Kaili Vesik, Muhammad Abdul-Mageed, and Miikka Silfverberg. 2020. One model to pronounce them all: Multilingual grapheme-to-phoneme conversion with a transformer ensemble. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 146–152, Online. Association for Computational Linguistics.

Yu-Chun Wang and Richard Tzong-Han Tsai. 2009. Rule-based Korean grapheme to phoneme conversion using sound patterns. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 2*, pages 843–850, Hong Kong. City University of Hong Kong.

Sonjia Waxmonsky and Sravana Reddy. 2012. G2P conversion of proper names using word origin information. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 367–371, Montréal, Canada. Association for Computational Linguistics.

Tomohiro Yamasaki. 2022. Grapheme-to-phoneme conversion for Thai using neural regression models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4251–4255, Seattle, United States. Association for Computational Linguistics.

Michiharu Yamashita, Hideki Awashima, and Hidekazu Oiwa. 2018. A comparison of entity matching methods between English and Japanese katakana. In *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 84–92, Brussels, Belgium. Association for Computational Linguistics.

Lars Yencken and Timothy Baldwin. 2005. Efficient grapheme-phoneme alignment for Japanese. In *Proceedings of the Australasian Language Technology Workshop 2005*, pages 143–151, Sydney, Australia.

A Background of Loanword in Japanese and the other Non-Latin Alphabet Writing System Languages

In languages with non-Latin-alphabet writing systems like Chinese or Korean, loanwords are often written using a mixture of their native script and the Latin characters, known as code-switching. However, in Japanese, loanwords are almost always transliterated into Katakana based on their pronunciation, causing various transliterations for the same word caused by individual differences and preferences, especially for new or not yet standardized NEs.

B The difficulty of Japanese G2P

The difficulty of Japanese G2P can be attributed not only to the complexity of its writing system, which employs three different systems (Hiragana, Katakana, Kanji), but also to the syllabic characteristics of Hiragana and Katakana. Unlike alphabetical writing systems such as Latin and Cyrillic, the basic unit of Hiragana and Katakana is a syllable, not a phoneme. A syllable is a kind of phonological unit, usually composed of a core (nucleus) vowel and potentially consonants before and/or after the core. In Hiragana and Katakana, characters representing different syllables are completely different from each other, even when they share the same vowel or consonant. For instance, the Katakana character for /ka/ is “カ”, while /ki/, which shares the same consonant, is represented by a totally different form: “キ”.

C Details of ParaNames entries

One of “PER”, “LOC”, and “ORG” is assigned to “type”, corresponding to “human”, “geographic region”, and “organization”, respectively, in the type in the original Wikidata. Table 3 shows the example of the data in ParaNames. Note that the number of language tags stored in “language” does not agree with the actual number of languages in the dataset, since a language may have multiple tags reflecting differences in writing systems, regions, and so on.[9]

D The number of notations per language tag in our dataset

The number of writing systems of original NE is 20 in total[10]. 15 languages have more pairs of NE-IPA for each than the approximately average number of pairs per language, one million. All of them use a Latin-based script except Russian and Chinese and belong to the Indo-European language family except Hungarian and Chinese. Table 5 shows the number of notations per each language tag.

[9] For example, there are two tags for Uzbek: “uz-cyrl” for Uzbek written in Cyrillic and “uz-latn” for Uzbek in the Latin script.

[10] Latin, Ge’ez, Arabic, Cyrillic, Bengali, Devanagari, Katakana, Hiragana, Khmer, Lao, Malayalam, Burmese, Oriya, Gurmukhi, Sinhalese, Tamil, Telugu, Thai, Simplified Chinese, and Traditional Chinese.

language tag: language name	number of notation	language tag: language name	number of notation
aa: Afar	28,894	ny: Chewa	29,309
am: Amharic	4,478	om: Oromo	30,961
ar: Arabic	830,648	or: Oriya	15,045
av: Avar	1,155	pa: Punjabi	18,733
az: Azerbaijani	115,014	pl: Polish	1,526,678
bn: Bengali	437,838	pt: Portuguese	373,237
ca: Catalan	3,057,109	pt-br: Portuguese (Brazil)	1,897,670
cs: Czech	1,283,030	rn: Rundi	28,017
de: German	4,177,379	ro: Romanian	894,080
de-at: German (Austria)	295,193	ru: Russian	1,333,970
de-ch: German (Switzerland)	31,433	rw: Kinyarwanda	30,374
de-formal: German (formal)	34	sg: Sango	27,867
en: English	13,715,761	si: Sinhala	19,001
en-ca: English (Canada)	424,497	sn: Shona	29,662
en-gb: English (UK)	141,742	so: Somali	30,748
es: Spanish	6,071,612	sq: Albanian	2,855,562
es-419: Spanish (Latin America)	1,271	sv: Swedish	2,733,009
es-formal: Spanish (formal)	506	sw: Swahili	294,945
fa: Persian	630,954	ta: Tamil	89,757
ff: Fulah	30,456	te: Telugu	52,395
fr: French	5,003,611	tg: Tajik	76,492
ha: Hausa	46,317	tg-cyrl: Tajik (Cyrillic)	566
hi: Hindi	74,366	tg-latn: Tajik (Latin)	29,674
hr: Croatian	426,170	th: Thai	91,844
ht: Haitian	88,589	ti: Tigrinya	288
hu: Hungarian	1,056,422	tk: Turkmen	28,707
hu-formal: Hungarian (formal)	2	tl: Tagalog	172,360
id: Indonesian	774,023	tr: Turkish	586,329
it: Italian	3,096,623	ug: Uighur	3,125
ja: Japanese	671,429	ug-arab: Uighur (Arabic)	113
jv: Javanese	151,768	uk: Ukrainian	654,641
kk: Kazakh	58,089	ur: Urdu	149,481
kk-cyrl: Kazakh (Cyrillic)	47,204	uz: Uzbek	192
kk-kz: Kazakh (Kazakhstan)	761	uz-cyrl: Uzbek (Cyrillic)	1
kk-latn: Kazakh (Latin)	74,802	uz-latn: Uzbek (Latin)	7
kk-tr: Kazakh (Turkey)	25,389	vi: Vietnamese	568,179
km: Khmer	3,241	xh: Xhosa	32,628
ky: Kyrgyz	44,936	yo: Yoruba	274,206
lo: Lao	1,408	zh: Chinese	965,861
mi: Maori	51,825	zh-cn: Chinese (simplified)	16,178
ml: Malayalam	93,555	zh-hans: Chinese (simplified)	255,533
mn: Mongolian	12,020	zh-hant: Chinese (traditional)	9,508
mr: Marathi	41,352	zh-hk: Chinese (Hong Kong)	5,065
ms: Malay	545,598	zh-mo: Chinese (Macao)	305
mt: Maltese	68,109	zh-my: Chinese (Mandarin in Malaysia)	6
my: Burmese	9,331	zh-sg: Chinese (Mandarin in Singapore)	104
nl: Dutch	9,586,075	zh-tw: Chinese (Mandarin in Taiwan)	9,518
nl-informal: Dutch (informal)	5		

Table 5: The number of notations per each language tag.

Author Index