Grapheme-to-phoneme conversion

Motivations (1/)

Speech technologies like automatic speech recognition and text-to-speech synthesis require mappings between written words and their pronunciations.

However, they are expensive to create and maintain, and free, large, high-quality dictionaries are only available for a small number of languages.

For open-vocabulary applications, these mappings must generalize to unseen words.

While it is often possible for a literate, linguistically-sophisticated native speaker to simply write out the rules, rule-based systems are brittle and difficult to maintain, and are often outperformed by machine learning techniques (e.g., van Esch et al. 2016).

Motivations (2/)

It is often possible for a literate, linguistically-sophisticated native speaker to simply write out the rules, but rule-based systems are brittle and difficult to maintain, and are often outperformed by machine learning techniques (e.g., van Esch et al. 2016).

Nearly all the prior evaluations have been conducted either on English or a few other highly-resourced alphabetic languages (e.g., Dutch, French, German, etc.).

This in turn is likely due to the lack of publicly available multilingual data...

Massively multilingual pronunciation modeling with WikiPron

Jackson L. Lee, Lucas F.E. Ashby[*], M. Elizabeth Garza[*], Yeonju Lee-Sikka[*], Sean Miller[*], Alan Wong[*], Arya D. McCarthy[†], and Kyle Gorman[*]

*: Graduate Center, City University of New York
†: Johns Hopkins University

LREC 2020

Wiktionary as a data source

Wiktionary (https://www.wiktionary.org/) is a free, collaboratively edited multilingual online dictionary, and some teams have previously used it for pronunciation data.

- Schlippe et al. (2010) extract Wiktionary pronunciation data for English, French, German, and Spanish. They report that this data is both abundant and improves automatic speech recognizer performance. However, they do not release any software or data.
- Deri and Knight (2016) release a collection of 650,000 word-pronunciation pairs extracted from Wiktionary. They too do not release the associated extraction software.

decimate

Contents [hide]

English [edit]

English Wikipedia has an article on:
decimation (Roman army)

Etymology [edit]

Borrowed from Latin *decimāre* (“to take or offer a tenth part”), from *decimus* (“tenth”).[1] As a noun, via Latin *decimatus* (“tithing area; tithing rights”).[2]

Pronunciation [edit]

- (*Received Pronunciation*) IPA(key): /ˈdɛ.sɪ.meɪt/
- Audio (UK) ▶ 0:00 MENU
- (*US*) enPR: dĕ.sə'māt", IPA(key): /ˈdɛ.sə.meɪt/

Verb [edit]

decimate (*third-person singular simple present* **decimates**, *present participle* **decimating**, *simple past and past participle* **decimated**)

1. (*archaic*) To kill one-tenth of a group, (*historical*, *specifically*) as a military punishment in the Roman army selected by lot, usually carried out by the surviving soldiers. [quotations ▼]
2. To destroy or remove one-tenth of anything. [quotations ▼]

WikiPron

WikiPron is an open-source library for mining pronunciations from Wiktionary.

While one can use it directly, users can take advantage of "the big scrape", a dynamic database of 3.1 million word/pronunciation pairs in 337 languages, dialects, and scripts, both living and dead, mined using WikiPron.

The big scrape is refreshed twice annually by our lab.

Pronunciation [edit]

- (*Portugal*) IPA(key): /ɡu.ˈɾi.lɐ/
- (*Brazil*) IPA(key): /ɡo.ˈɾi.lɐ/
- Hyphenation: go·ri·la

Pronunciation [edit]

- IPA(key): (*most of Spain and Latin America*) /ʝaˈmaɾ/, [ɟ͡ʝaˈmaɾ]
- IPA(key): (*rural northern Spain, Andes Mountains*) /ʎaˈmaɾ/, [ʎaˈmaɾ]
- IPA(key): (*Buenos Aires and environs*) /ʃaˈmaɾ/, [ʃaˈmaɾ]
- IPA(key): (*elsewhere in Argentina and Uruguay*) /ʒaˈmaɾ/, [ʒaˈmaɾ]

Scraping features

- Narrow ([phonetic]) versus broad (/phonemic/) transcription
- Whether or not transcriptions should include:
  - stress markings
  - syllable boundaries
  - tones
- Whether to segment the transcriptions (e.g., kʰæt → kʰ æ t)
- Whether only entries from a specific (inputted) dialect(s) should be included
- Whether or not to case-fold the headword

ISO 639-2 Code	ISO 639 Language Name	Wiktionary Language Name	Script	Dialect	Filtered	Narrow/Broad	Case-folding	# of entries
aar	Afar	Afar	Latin		False	Broad	True	715
acw	Hijazi Arabic	Hijazi Arabic	Arabic		False	Broad	False	1,090
acw	Hijazi Arabic	Hijazi Arabic	Arabic		False	Narrow	False	167
ady	Adygei; Adyghe	Adyghe	Cyrillic		False	Narrow	True	5,123
ady	Adygei; Adyghe	Adyghe	Cyrillic		True	Narrow	True	4,895
afb	Gulf Arabic	Gulf Arabic	Arabic		False	Broad	False	528
afr	Afrikaans	Afrikaans	Latin		False	Broad	True	1,685
afr	Afrikaans	Afrikaans	Latin		True	Broad	True	1,659
afr	Afrikaans	Afrikaans	Latin		False	Narrow	True	121
ajp	South Levantine Arabic	South Levantine Arabic	Arabic		False	Broad	False	155
alb	Albanian	Albanian	Latin		False	Broad	True	1,450
alb	Albanian	Albanian	Latin		False	Narrow	True	823
ale	Aleut	Aleut	Latin		False	Broad	True	104
ang	Old English (ca. 450-1100)	Old English	Latin		False	Broad	True	8,854
ang	Old English (ca. 450-1100)	Old English	Latin		False	Narrow	True	4,341
aot	Atong (India)	Atong (India)	Latin		False	Broad	True	140
apw	Western Apache	Western Apache	Latin		False	Narrow	True	158
ara	Arabic	Arabic	Arabic		False	Broad	False	7,279
arc	Imperial Aramaic (700-300 BCE); Official Aramaic (700-300 BCE)	Aramaic	Hebrew		False	Broad	False	1,156
arm	Armenian	Armenian	Armenian	Eastern Armenian, standard	False	Narrow	True	14,182
arm	Armenian	Armenian	Armenian	Eastern Armenian, standard	True	Narrow	True	14,177
				Western				

Features of the pronunciation dictionary library/database

- By default all words are case-folded and all transcriptions are segmented with stress and syllable boundaries removed.
- Beyond this multiple post-processing steps are applied to the data:
    - Languages using multiple scripts are split into separate TSVs.
    - Alternate filtered TSVs are generated for specific languages.

105	nłt’éégo	ṇ ɬ t’ é: k ò
106	ooro	o: r o
107	piishi	pʰ ɪ: t ʃ ɪ
108	pish	pʰ ɪ̀ ʃ
109	shash	ʃ a ʃ

Filtering the results of the scrape using phones lists

English pronunciation for the word ‘Bach’:

Pronunciation [edit]

- (*Received Pronunciation*) IPA(key): /baːx/, /baːk/

Handwritten lists of permitted phones allow us to filter pronunciations for languages (and dialects of languages).

```
b
d
f
g
h
j
k
l
m
n
r
p
s
t
w
a
a:
e
e:
i
i:
o
o:
u
u:
u̯  # Offglide of the <au, eu> diphthongs.
e̯  # Offglide of the <ae, oe> diphthongs.
h
kʷ
gʷ
# And in Greek borrowings only:
pʰ
tʰ
kʰ
y
y:
z
```

Challenges

Most languages on Wiktionary use the same underlying HTML structure for their entries. Those that don’t require bespoke extraction functions. Changes to Wiktionary or the underlying HTML of particular languages on Wiktionary often leads to scraping failure.

A majority of development time on WikiPron has been dedicated to handling differences in the HTML underlying entries in specific languages.

Reliability engineering

WikiPron uses extensive continuous integration testing (build, testing with `pytest`, static type checking with `mypy`, linting with `flake8`, reflowing with `black`) via CircleCI and GitHub's webhook integration.

WikiPron workflows (like the big scrape) produce human-readable tables and TSV summaries as a side-effect.

Ongoing development

- Phonelist development
- Addition of a ‘subdialect’ flag or some method for handling dialects within dialects
- Testing for large-scale changes to the scraping module
- Prospective upstream improvements to Wiktionary itself

The SIGMORPHON shared tasks on grapheme-to-phoneme conversion

The SIGMORPHON 2020 Shared Task on Multilingual Grapheme-to-Phoneme Conversion

Kyle Gorman[*], Lucas F.E. Ashby[*], Aaron Goyzueta[*],
Arya D. McCarthy[†], Shijie Wu[†], and Daniel You[‡]

*: Graduate Center, City University of New York
†: Johns Hopkins University
‡: Jericho High School

Results of the Second SIGMORPHON shared task on multilingual grapheme-to-phoneme conversion

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*, Winnie Yan*

*Graduate Center, City University of New York
†University of Zurich

Methods

- Words with multiple pronunciations are excluded:
    - Some represent real "variants".
    - Others are homographs, for which see Gorman et al. 2018, Seale 2021, etc.
- Words are sampled according to their frequency in the Wortschatz (Goldhahn et al. 2012) frequency norms if available, or uniformly if not.
- The data is randomly split into 80% training data, 10% development data, and 10% test data. (Splitting is "lexeme-aware" thanks to UniMorph.)
- Pronunciations are segmented using `segments` (Moran & Cysouw 2018).
- Systems are ranked according to macro-averaged *word error rate* (WER).

Task design

- 2020: 10 basic languages, 5 surprise languages, 4,500 examples each
- 2021:
    - High-resource subtask: 40,000 words of American English, all external resources permitted (except Wiktionary pronunciation mining tools)
    - Medium-resource subtask: 10,000 words, 10 languages, UniMorph paradigms permitted
    - Low-resource subtask: 1,000 words, 10 languages, no external resources permitted

New this year

- QA for the WikiPron data backend, including:
  - *phonelist filtration*
  - *automated script detection*
  - manual post-extraction fixes for English, Bulgarian, Maltese (Latin), and Welsh
- New subtasks (all 80%/10%/10% split):
  - high-resource (with arbitrary third-party resources): 1 language x 41,000 examples
  - medium-resource (with UniMorph paradigms—though nobody used them): 10 languages x 10,000 examples
  - low-resource (with no third-party resources): 10 languages x 1,000 examples
- Semi-automated error analysis (more on that in a second)

Language	ISO 639-2	Example training data pair	
Armenian	`arm`	մեծաքանակ	m ɛ t͡s ɑ kʰ ɑ n ɑ k
Bulgarian	`bul`	североизток	s ɛ v ɛ r o i s t o k
French	`fre`	hébergement	e b ɛ ʁ ʒ ə m ɑ̃
Georgian	`geo`	ფორმიანი	pʰ ɔ r m ɪ ɑ n ɪ
Modern Greek	`gre`	καθισμένες	k a θ i z m e n e s
Hindi	`hin`	कैलकुलेटर	k ɛː l k ʊ l eː ʈ ə ɾ
Hungarian	`hun`	csendőrök	t͡ʃ ɛ n d øː r ø k
Icelandic	`hin`	þýskaland	θ i s k a l a n t
Korean	`kor`	말레이시아	m a̠ ɭ ɭ e̞ i ɕʰ i a̠
Lithuanian	`lit`	galinčiais	ɡ aː lʲ ɪ nʲ tʲ ʃʲ ɛ j s
Adyghe	`ady`	бзыукъолэн	b z ə w qʷ a l a n
Dutch	`dut`	aanduiding	aː n d œ y̯ d ɪ ŋ
Japanese hiragana	`jpn`	どちらさま	d o̞ t͡ɕ i ɾ a̠ s a̠ m a̠
Romanian	`rum`	bineînțeles	b i n e ɨ n t s e l e s
Vietnamese	`vie`	duyên phận	z w i ə n ˧˧ f ə n ˨˩ ʔ

Table 1: Languages, language codes, and example training data pairs for the shared task.

Armenian (Eastern)	`arm_e`	համադրություն	h ɑ m ɑ d ə ɾ u tʰ j u n
Bulgarian	`bul`	обоснованият	o̞ b o̞ s n o v a n i j ə t̪
Dutch	`dut`	konijn	k oː n ɛ i̯ n
French	`fre`	joindre	ʒ w ɛ̃ d ʁ̞̃
Georgian	`geo`	მოუქნელად	m ɔ u kʰ n ɛ l ɑ d
Serbo-Croatian (Latin)	`hbs_latn`	opadati	o p ǎː d a t i
Hungarian	`hun`	lobog	l o b o g
Japanese (Hiragana)	`jpn_hira`	ぜんたいしゅぎ	d͡z ẽ̞ n t a̠ i ɕ ɨᵝ gʲ i
Korean	`kor`	쇠가마우지	sʰ w e̞ g a̠ m a̠ u d͡ʑ i
Vietnamese (Hanoi)	`vie_hanoi`	ngừng bắn	ŋ ɨ ŋ ˨˩ ʔ ɓ a n ˧˥

Table 1: The ten languages in the medium-resource subtask with language codes and example training data pairs.

Adyghe	`ady`	кӀэшӀыхьан	t͡ʃʼ a ʃʼ ə ħ aː n
Greek	`gre`	λέγεται	l e ʝ e t e
Icelandic	`ice`	maður	m aː ð ʏ r
Italian	`ita`	marito	m a r i t o
Khmer	`khm`	ប្រហារ	p r ɑ h aː
Latvian	`lav`	mīksts	m îː k s t s
Maltese (Latin)	`mlt_latn`	minna	m ɪ n n a
Romanian	`rum`	ierburi	j e r b u rʲ
Slovenian	`slv`	oprostite	ɔ p r ɔ s t íː t ɛ
Welsh (Southwest)	`wel_sw`	gorff	g ɔ r f

Table 2: The ten languages in the low-resource subtask with language codes and example training data pairs.

Baselines

- 2020:
  - A pair n-gram model implemented using OpenGrm (Novak et al. 2016)
  - An LSTM attentive encoder-decoder sequence-to-sequence model (Luong et al. 2015)
  - A transformer encoder-decoder sequence-to-sequence model (Vaswani et al. 2017)
- 2021:
  - An imitation learning-based neural transducer (Makarov & Clematide 2018)

Baselines

- 2020:
  - A pair n-gram model implemented using OpenGrm (Novak et al. 2016)
  - **An LSTM attentive encoder-decoder sequence-to-sequence model (Luong et al. 2015)**
  - A transformer encoder-decoder sequence-to-sequence model (Vaswani et al. 2017)
- 2021:
  - An imitation learning-based neural transducer (Makarov & Clematide 2018)

Submissions

- 2020: 23 submissions from 9 teams; **IMS** achieves 3% absolute WER reduction
- 2021: 13 submissions from 4 teams:
  - High-resource subtask: **Dialpad** achieves 4% absolute WER reduction
  - Medium-resource subtask: no team beats the baseline
  - Low-resource subtask: **UBC** achieves 1% absolute WER reduction

2021 error analysis

Two methods were used:

- An automated accounting of the most common errors per language across all submissions (after Makarov & Clematide 2020)
- An automated sorting of errors into
    - errors consistent with a hand-written finite-state covering grammar (*model deficiencies*, usually due to inherent ambiguity in the orthography) vs.
    - errors not consistent with the covering grammar (*coverage deficiencies*, usually indicating inconsistencies in the gold data itself).

eng_us	ɪ	ə	113	ɑ	ɔ	112	_	ʊ•	96	_	ɪ•	85	ɪ	i	76
arm_e	_	ə•	16	ə•	_	10	tʰ	d	6	d	tʰ	6	j•	_	3
bul	ɛ•	d͡	32	a	ə	31	ə	ɤ	30	_	◌̞	27	ə	a	25
dut	ə	eː	10	_	ː	10	ə	ɛ	9	eː	ə	8	z	s	8
fre	a	ɑ	6	_	•s	5	ɔ	o	5	e	ɛ•ʁ	3	_	•t	3
geo															
hbs_latn	_	ː	85	ː	_	76	_	◌̆	55	◌̆	◌̂	53	◌̆	_	52
hun	_	ː	6	h	ɦ	3	ʃ	s	2	ː	_	2			
jpn_hira	_	◌̥	20	_	◌̊	11	_	d͡	4	ː	•ɰ̟ᵝ	3	h	ɰᵝ	3
kor	_	ː	73	ː	_	28	ʌ̹	ɘː	23	ʰ	◌̤	9	ɘː	ʌ̹	6
vie_hanoi	_	w•	3	_	˧˩	3	_	w•ŋ͡m•	2	◌͡ɕ	•ɹ	2	_	ʔ•	2
ady	ʼ	_	3	ː	_	3	ʃ	ʂ	3	ə•	_	2	a	ə	2
gre	ɾ	r	8	r	ɾ	3	i	ʝ	3	m•	_	2	ɣ	ɡ	2
ice	ː	_	2	◌̥	_	2	_	ː	2						
ita	o	ɔ	6	e	ɛ	5	j	i	3	◌͡	•	2	ɔ	o	2
khm	aː	i•ə	3	_	ʰ	3	_	•ɑː	2	ĕ	ɔ	2	ɑ	a	2
lav	◌̄	◌̂	11	_	◌̂	10	◌̀	_	9	◌̄	_	7	_	◌̀	4
mlt_latn	_	ː	5	_	ɪ•	2	ɐ	a	2	b	p	2	a	ɐ	2
rum	◌͡	•	2												
slv	◌́	◌̀	7	◌̀ː	_	6	◌́ː	_	6	_	◌́ː	5	ɛ	éː	4
wel_sw	ɪ	iː	3	ɪ	i̯	2	_	ɛ•	2						

Table 7: The five most frequent error types, represented by the hypothesis string, gold string, and count, for each language; • indicates whitespace and _ the empty string.

	Baseline		CLUZH-5	
	WER	MDR	WER	MDR
bul	18.3	17.6	19.2	19.0
fre	8.5	7.5	7.5	6.8
jpn_hira	5.2	4.4	5.3	4.5

Table 9: WER and model deficiency rate (MDR) for three languages from the medium-resource subtask.

	Baseline		AZ		CLUZH-1		UBC-2	
	WER	MDR	WER	MDR	WER	MDR	WER	MDR
ady	22	22	30	23	24	21	22	22
gre	21	18	23	19	20	17	22	21
ice	12	9	22	17	10	7	11	5
ita	19	15	25	19	23	16	22	19

Table 10: WER and model deficiency rate (MDR) for four languages from the low-resource subtask.

Discussion

- Substantial across-the-board improvement in performance from 2020 to 2021:
    - Better modeling
    - More data in medium-resource condition
    - Better quality control (Georgian is at ceiling!)
- Large gap between higher- and lower-resource subtasks remains:
    - Baseline achieves WER of 10.6 in medium-resource scenario, but
    - just a WER of 25.1 in the low-resource scenario
    - Model data efficiency is not sufficient to generalize well with just 800 examples
- Error analysis suggests that much of the residual error is due to *inherent* orthographic ambiguity
- No participants have as of yet experimented with morphological decompositions, features, or lemmata.

Resources:

https://unimorph.github.io/
https://github.com/CUNY-CL/wikipron
https://github.com/sigmorphon/2020
https://github.com/sigmorphon/2021-task1