Orthography versus IPA: Why We Need Both

Anjali Kantharuban and David R. Mortensen

March 12, 2024

Sometimes natural language processing practitioners ask "Why would we need the International Phonetic Alphabet? Don't ordinary alphabets represent sound? If we need access to speech sounds, we can recover it from a character-level model." But there are three reasons that IPA is needed in addition to orthography:

1. The mapping from phonemes to orthography is ambiguous.
2. The mapping from orthography to phonemes is ambiguous.
3. The mappings between phonemes and orthography are language (and even dialect) dependent.

One Phonemic Word, Many Orthographic Words

English is not unique in having sets like the following:

(1) a. ⟨maze⟩
b. ⟨maize⟩
c. ⟨Mays⟩

All of these correspond to phonemic /meiz/. In isolation, there is no way of knowing whether the intended lexeme is the labyrinth, the grain, or the months. Further more, there is now way of knowing whether it is spelled with ⟨aze⟩, ⟨aize⟩, or ⟨ays⟩. In actual writing, we disambiguate according to context (as with the nouns ⟨principle⟩ and ⟨principal⟩ or the adjectives ⟨discreet⟩ and ⟨discrete⟩).

French is full of this kind of ambiguity (know as HOMOPHONY. Compare the following:

(2) a. ⟨pair⟩ 'peer'
b. ⟨paire⟩ 'pair'
c. ⟨père⟩ 'father'

However, Chinese puts all other levels of sound-to-symbol ambiguity to shame. Consider the following words that are pronounced /puˇ/:

(3) a. 不 'negative'
b. 埔 'port; wharf; pier'
c. 埗 'wharf; dock; jetty'
d. 埠 'wharf; port; pier'
e. 布 'cloth'
f. 怖 'terror'
g. 捗 'make progress'
h. 步 'step'

i. 瓶 'kind of vase'
j. 箘 'sieve-like-utensil'
k. 簿 'book'
l. 部 'ministry'
m. 钚 'plutonium'

This kind of ambiguity is not a major obstacle with regards to extracting representations of sound, but it makes the conversion of representations of pronunciation to orthography more challenging. This task is not possible without the use of some type of language model.

One Orthograpic Word, Many Phonemic Words

A more troublesome kind of ambiguity is that in which there are many phonemic words for one orthographic words. In other words, the orthography loses information. This is also very common in English. Consider the following examples:

(4) a. advocate (noun) — advocate (verb)
b. affiliate (none) — affiliate (verb)
c. alternate (verb) — alternate (noun)
d. bass (fish) — bass (musical instrument, vocal part)
e. bow (noun) — bow (verb)
f. house (noun) — house (verb)
g. live (adjective) — live (verb)

The upper bound in terms of this kind of ambiguity is probably manifest in languages written with ABJADS[1] like Arabic and Hebrew. In these writing systems, short vowels are not represented at all. Therefore, any words that differ only in their short vowels will be written the same. Consider the following:

(5) a. بتك *kataba* 'wrote'
b. بتك *kutiba* 'was written'

Although these *can* be written with special diacritics that represent the short vowels, this is typically only done with religious texts (where a small ambiguity could have large consequences in terms of religious teaching and practice). In practice, these word forms are written the same, even though this difference is sometimes not predictable from the larger context.

If our goal is to extract phonemic (IPA) transcriptions from orthographic strings, this phenomenon presents a challenge (since it requires recovering information that is latent within a sentence or discourse but is not present in the spelling of individual words).

[1] An abjad is a writing system in which consonants an sometimes long vowels (but no short vowels) are represented. Otherwise, an abjad is like an ALPHABET.

Sound-Symbol Mapping is Language Specific

Varieties of Chinese other than Mandarin are now not typically written (the exception being Cantonese), but prior to the early 20th century, there was no single standardized variety and everybody wrote in an archaic variety and read it aloud in the local variety (of which there are many). These varieties are, in fact, very different from one another in pronunciation and are treated by linguists as separate languages, but they were written with a single orthography. As show in Table 1, the pronunciation of a single character varied tremendously, even within a closely related group of language varieties like Chinese. This is probably the most extreme version of a generalization:

Orthography	Cantonese	Gan	Hakka	Jin	Mandarin	Hokkien	Wu	Xiang
犬	hyːn˧	t͡ɕʰy̯ən˨˩	kʰi̯en˩	t͡ɕʰy̯e˨	t͡ɕʰy̯ɛn˩	kʰi̯ɛn˩˧	t͡ɕʰy̯ø˧	t͡ɕʰy̯e̞˩˧
月	jyːt˨	ɲy̯ət˨	ɲi̯et˥	y̯ə?˨	y̯ɛ˩˧	ge?˥	ɦy̯ɪ?˨	y̯e̞
人	jɐn˨	ɲin˧	ɲin˨	zə̃ŋ˨	z̩ən˧	d͡ʑin˧	ɲɪŋ˧	z̩ən

Table 1: Some mappings from Han characters to Chinese varieties.

sound-symbol mapping is specific to particular languages and language varieties.

If anything, the problem is more pronounced across languages. The Latin alphabet is widely thought of as being somewhat transparent and orthographies based on this script throughout the world share a common origin, but just in the Epitran G2P system, Latin-script languages use ⟨x⟩ to represent the following sounds:

- ks
- ɕ
- ħ
- s
- x
- χ
- ʃ
- z
- kǁ
- ɖ
- t’
- ǁ

When it is useful to compare the sounds of language cross-linguistically, orthography is clearly not up to the job.

However, why would this be useful? Why *would* you need a language-neutral way of representing the sounds of languages:

(6) a. Natural language processing is mostly about written words, divorced from pronunciation

b. For speech processing, acoustic signals exist which are both language-neural and easy to obtain

There are at least three reasons that IPA, rather than orthography, is useful for speech and language scientists:

(7) a. The scientific study of speech and language requires making reference to sounds in such a way that they can be compared cross-linguistically

b. We need a way of distinguishing between normative and non-normative pronunciation (as in speech therapy or second language teaching and assessment) and this is possible only if we have a systematic way of indicating non-normative speech

c. Similarities in the pronunciation of words can provide useful signals in NLP tasks (for example, NER, QA, MT, and Entity Linking)[2]

In future lectures, we will discuss how to transduce orthographic representations (which are commonly available) into phonemic representations. This task is called grapheme-to-phoneme transduction or G2P. In future lectures, we will learn about rule-based and data-driven approaches to G2P and will learn how to implement rule-based G2P for a new language using Epitran[3].

References

Akash Bharadwaj, David Mortensen, Chris Dyer, and Jaime Carbonell. Phonologically aware neural model for named entity recognition in low resource transfer settings. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1462–1472, Austin, Texas, November 2016. Association for Computational Linguistics. DOI: 10.18653/v1/D16-1153. URL https://aclanthology.org/D16-1153.

Aditi Chaudhary, Chunting Zhou, Lori Levin, Graham Neubig, David R. Mortensen, and Jaime Carbonell. Adapting word embeddings to new languages with morphological and phonological subword representations. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3285–3295, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. DOI: 10.18653/v1/D18-1366. URL https://aclanthology.org/D18-1366.

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. Epitran: Precision G2P for many languages. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1429.

[2] Akash Bharadwaj, David Mortensen, Chris Dyer, and Jaime Carbonell. Phonologically aware neural model for named entity recognition in low resource transfer settings. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1462–1472, Austin, Texas, November 2016. Association for Computational Linguistics. DOI: 10.18653/v1/D16-1153. URL https://aclanthology.org/D16-1153; Akash Bharadwaj, David Mortensen, Chris Dyer, and Jaime Carbonell. Phonologically aware neural model for named entity recognition in low resource transfer settings. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1462–1472, Austin, Texas, November 2016. Association for Computational Linguistics. DOI: 10.18653/v1/D16-1153. URL https://aclanthology.org/D16-1153; and Aditi Chaudhary, Chunting Zhou, Lori Levin, Graham Neubig, David R. Mortensen, and Jaime Carbonell. Adapting word embeddings to new languages with morphological and phonological subword representations. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3285–3295, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. DOI: 10.18653/v1/D18-1366. URL https://aclanthology.org/D18-1366

[3] David R. Mortensen, Siddharth Dalmia, and Patrick Littell. Epitran: Precision G2P for many languages. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1429