The XPF Corpus: Rule-based grapheme to phoneme translation schemes for hundreds of languages

Uriel Cohen Priva

Brown University

We introduce the XPF Corpus, a grapheme-to-phoneme (G2P) engine and a collection of G2P translation schemes. This resource is intended for use in linguistics, language preservation, and language revitalization.

Over the past two decades, there has been an explosion of word-level data availability, even for low-resource languages (e.g., Scannell, 2007). However, the ability to utilize such resources in phonological and phonetic analysis has been significantly hindered by the challenge of trans-lating written language into phonological and phonemic annotations (McCarthy et al., 2023). Grapheme-to-phoneme (G2P) applications aim to bridge this gap, offering automated and semi-automated G2P engines (e.g., Hammond, 2023; McAuliffe et al., 2017). Additionally, a few G2P engines are hand-written (Epitran, Mortensen et al., 2018).

However, the engineering goals in G2P systems often diverge from those of phoneticians and phonologists, impacting both the output and the scope of most existing G2P engines. First, from an engineering perspective, it makes sense to prioritize fault tolerance: if an input word deviates from the standard spelling convention of a language, the system should still attempt to translate it. In contrast, a linguistic analysis should flag such non-standard words, which may follow from code-switching. Second, languages with many speakers and ample resources facilitate the training of semi-automatic G2P engines, but linguists often focus on extremely low-resource languages spoken by small communities. Finally, most existing G2P engines cater to experts and lack user-friendly features for end-users and local communities to create and test G2P rules.

The Cross-linguistics Phonological Frequencies (XPF) Corpus aims to address this gap for languages that have a deterministic grapheme-to-phoneme (G2P) translation scheme (Cohen Priva et al., 2021). The implementation flags cases where a grapheme cannot be translated. Each rule set is accompanied by a concise phonological sketch that describes the underlying assumptions of the translation scheme, and a list of reference word translation to showcase and test the rule set. Additionally, the corpus includes a reference Python implementation, along with a JavaScript implementation that can be used online, as in https://cohenpr-xpf.github.io/XPF/. Currently, the corpus provides resources for over two hundred languages, many of which are severely under-documented (see the table and figure on the next page). To the best of our knowledge, it supports more languages than any other hand-written collection of G2P rule sets.

Rule sets are based on a small number of rules types, which together enable the translation of most languages using relatively few rules. This approach makes rule sets more manageable, at the expense of requiring rule set authors to use regular expressions. The median number of rules in XPF rule sets is 28 [cf. 47 in Epitran]. In addition to grapheme-to-phoneme (G2P) rules, the corpus extends G2P functionality in several ways. Word translation rules make it possible to combine G2P rules with existing dictionaries. Preprocessing rules address less common upper and lower case schemes, such as in Azerbaijani, in which lowercase *I* and *İ* are *ı* and *i*, respectively. Finally, phoneme-to-phoneme translation rules use regular expressions to e.g. translate doubled phonemes into phoneme and length symbols.

The corpus has already been used by several researchers, but it has not yet been presented in professional venues. Our goal is to instruct others on how to adapt the corpus to their specific requirements and effectively utilize existing G2P rule sets.

Language family	Number of languages in language family
Arawakan	9
Austronesian	28
Indo-European	24
Mayan	12
Niger-Congo	7
Trans-New Guinea	24
Turkic	13
Other	84
Total	201

Table 1: The distribution of languages in the XPF Corpus. *Other* stands for language isolates, language groups that are represented by fewer than 5 languages, and all creoles.

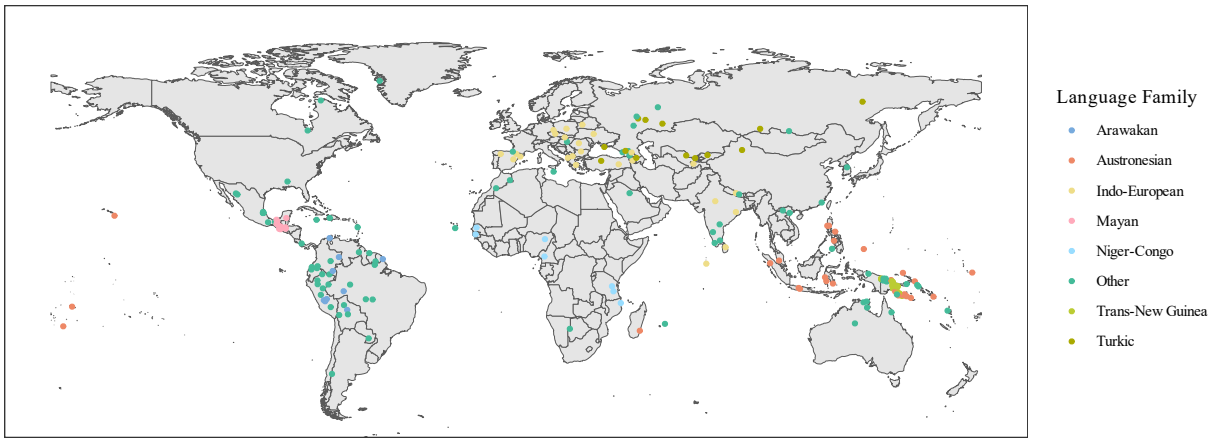

Figure 1: World Map of XPF Corpus Languages

References

Cohen Priva, U., Strand, E., Yang, S., Mizgerd, W., Creighton, A., Bai, J., Mathew, R., Shao, A., Schuster, J., & Wiepert, D. (2021). *The cross-linguistic phonological frequencies (XPF) corpus manual*.

Hammond, M. (2023). Low-resource grapheme-to-phoneme mapping with phonetically-conditioned transfer. In *SIGMORPHON proceedings*.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proc. Interspeech 2017*.

McCarthy, A. D., Lee, J. L., DeLucia, A., Bartley, T., Agarwal, M., Ashby, L. F. E., Del Signore, L., Gibson, C., Raff, R., & Wu, W. (2023). The SIGMORPHON 2022 shared task on cross-lingual and low-resource grapheme-to-phoneme conversion. In *SIGMORPHON proceedings*.

Mortensen, D. R., Dalmia, S., & Littell, P. (2018). Epitran: Precision G2P for many languages. In *LREC proceedings*.

Scannell, K. P. (2007). The Crúbadán project: Corpus building for under-resourced languages. In *Building and exploring web corpora: Proceedings of the 3rd web as corpus workshop* (Vol. 4).