On Pronunciations in Wiktionary: Extraction and Experiments on Multilingual Syllabification and Stress Prediction

Winston Wu and **David Yarowsky**
Center for Language and Speech Processing
Johns Hopkins University
{wswu,yarowsky}@jhu.edu

Abstract

We constructed parsers for five non-English editions of Wiktionary, which combined with pronunciations from the English edition, comprises over 5.3 million IPA pronunciations, the largest pronunciation lexicon of its kind. This dataset is a unique comparable corpus of IPA pronunciations annotated from multiple sources. We analyze the dataset, noting the presence of machine-generated pronunciations. We develop a novel visualization method to quantify syllabification. We experiment on the new combined task of multilingual IPA syllabification and stress prediction, finding that training a massively multilingual neural sequence-to-sequence model with copy attention can improve performance on both high- and low-resource languages, and multi-task training on stress prediction helps with syllabification.

1 Introduction

Wiktionary[1] is a free online multilingual dictionary containing a plethora of interesting information. In this paper, we focus on the pronunciation annotations in Wiktionary, which are relatively understudied. For any given word, Wiktionary may include data for IPA (both phonetic and phonemic), hyphenation, dialectical variation, and even audio files of speakers pronouncing the words. These types of data have been shown to be useful for tasks such as grapheme-to-phoneme transduction, e.g. in recent SIGMORPHON shared tasks (Gorman et al., 2020). There are many existing parsing efforts that have extracted pronunciation information from Wiktionary. Recent extractions of data from Wiktionary focus on obtaining high-quality pronunciations from a *single* edition of Wiktionary, usually the English edition (e.g. Wu and Yarowsky, 2020a; Sajous et al., 2020; Lee et al., 2020). However, substantial increases in data can be obtained by parsing other editions of Wiktionary, which have been shown to be helpful for downstream tasks. For example, Schlippe et al. (2010) extract pronunciations from the English, French, German, and Spanish editions, and **?** extract pronunciations from the English, German, Greek, Japanese, Korean, and Russian editions.

In this paper, we build upon Yawipa (Wu and Yarowsky, 2020a,b), a recent Wiktionary parsing framework. Targeting the larger Wiktionaries for increased coverage and those not dealt with in previous work, we construct new pronunciation parsers for the French, Spanish, Malagasy, Italian, and Greek editions of Wiktionary. Combined with pronunciations from the English Wiktionary, this totals to over 5.3 million words, which to our knowledge is the largest pronunciation lexicon to date and also a unique comparable corpora of pronunciations. In Section 2, we show that our extracted pronunciations are a substantial increase in data, covering numerous pronunciations not in the English Wiktionary. This is especially beneficial for low-resource languages. In Section 3, we analyze this data and find that a small portion of these pronunciations may be low-quality and computer-generated. In Section 4, we present a novel visualization technique for analyzing the use of stress in IPA pronunciations. In Section 5, we experiment on the combined task of massively multilingual syllabification and stress detection. Our neural sequence-to-sequence model with copy attention outperforms a sequence labeling baseline, especially in very low-resource scenarios, underscoring the contributions of additional languages to the task. In addition we find that a multi-task approach of predicting both stress and syllabification can improve the performance on syllabification alone.

[1] www.wiktionary.org

Proceedings of 14th Workshop on Building and Using Comparable Corpora in conjunction with RANLP-2021, pages 68–74, held online, 6 September 2021.
https://doi.org/10.26615/978-954-452-076-2_009

Figure 1: Screenshot of the English Wiktionary's pronunciation information for the French word *chien*.

2 Wiktionary Pronunciation Extraction

As a multilingual resource, Wiktionary exists as a set of numerous *editions*. That is, the English Wiktionary is written in English by and for English speakers, while the French Wiktionary is written in French by and for French speakers. Any edition can contain entries for words in any language. For example, Figure 1 shows a screenshot of the English Wiktionary's pronunciation information for the French word *chien*. We use the terms <*lang*> *edition* and <*lang*> *Wiktionary* interchangeably.

Why parse other editions of Wiktionary? Speakers of different languages have different priorities when annotating data. One can assume that an editor of the Spanish Wiktionary is more likely to provide pronunciations for Spanish words before working on English words. Our effort at extracting a new dataset of pronunciations from 6 different editions of Wiktionary resulted in a total of over 5,3 million *unique* IPA pronunciations across 2,177 languages. Note that because the data comes from multiple editions, a word may have multiple annotated pronunciations, making our dataset an interesting comparable corpora. Figure 2 shows the 16 languages with the most data in this dataset, along with the contribution of each edition of Wiktionary from which we parsed and extracted IPA pronunciations.

We draw several insights from Figure 2. First, the inclusion of pronunciations from non-English Wiktionaries represents substantial gains. Though the English edition is the largest Wiktionary by number of entries,[2] the French edition contains a huge number of pronunciations for French words, dwarfing other editions that we parsed. The French Wiktionary also supplies the entirety of the pronunciations for Northern Sami words (`se`, spoken in Norway, Sweden, and Finland), most of the available pronunciations for Esperanto (`eo`) and Italian

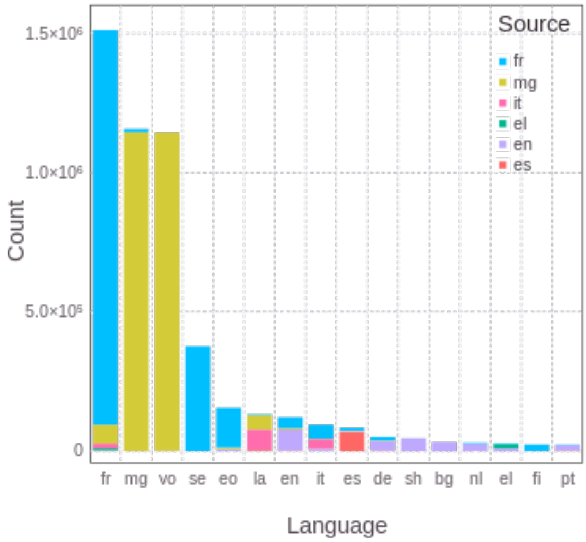

Figure 2: The top 16 languages in terms of number of pronunciations, with contributions from multiple editions of Wiktionary.

(`it`) words, and also words in 1,198 other low-resource languages not shown in the long tail of Figure 2. In contrast, the English edition (the second largest supplier) is the sole supplier of pronunciations in 416 languages.

Parsing Implementation The Yawipa framework (Wu and Yarowsky, 2020a) extracts data from the XML dump of Wiktionary.[3] Every entry is encoded in MediaWiki markup, which is similar to Markdown but includes special *templates* (enclosed in double braces) which programmatically generates HTML that we see when we visit the Wiktionary website. For example, in the English wiktionary, the entry for the French word *chien* contains the following markup (rendered in Figure 1):

```
===Pronunciation===
{{fr-IPA}}
{{audio|fr|Fr-chien.ogg|audio}}
{{rhymes|fr|jɛ̃}}
```

These three templates generate the three bullet points in Figure 1. Note that the `{{fr-IPA}}` template generates the IPA pronunciation, so the IPA itself does not exist in the English Wiktionary dump. Thus, we can only extract the IPA from the French edition, below, highlighting the need to parse multiple Wiktionary editions for multiple sources of pronunciations.

[2] https://meta.wikimedia.org/wiki/Wiktionary

[3] https://dumps.wikimedia.org/enwiktionary/latest/XXwiktionary-latest-pages-articles.xml.bz2, where XX is replaced with a two-letter ISO 639-1 code.

```
=== {{S|nom|fr}} ===
{{fr-rég|ʃjɛ̃}}
```

Above is the French Wiktionary's pronunciation for the word *chien*. A template (`fr-rég`) is also used, but the IPA is extractable from the markup. Each edition of Wiktionary has its own conventions on formatting and templates, thus requiring a separate parser specifically for that edition. For implementation details, please see the repository https://github.com/wswu/yawipa.

3 Analysis of the Dataset

For high-resource languages, the home language edition (e.g. English edition for the English language) usually supplies the most pronunciations, but this is not always the case (e.g. the French Wiktionary provides more Italian pronunciations than the Italian edition). In terms of amount of data, two languages are outliers: Malagasy (`mg`, an Austronesian language spoken in Madagascar) and Volapük (`vo`, a constructed language). As relatively less spoken languages, these languages have a disproportionately large amount of data. Why is this so?

The data for these two languages come from the Malagasy edition, which we parsed because of its high ranking in the List of Wiktionaries.[4] Both Malagasy and Volapük are inflected languages[5] whose IPA pronunciations seem to be entirely computer-generated using a regular transduction process from orthography to IPA, which was exploited to create a large set of pronunciations for these two languages.

We also find that some Latin pronunciations may be machine-generated. For example, the Malagasy edition supplies /kontabulawit/ as the pronunciation for the Latin *contabulavit* and [d̪ẽːonstrat] for *demonstrat*. These pronunciations lack stress and syllable markings, and in the case of *demonstrat*, do not agree with established pronunciations of Latin. thus leading us to believe that these were machine-generated pronunciations. In contrast, the English edition contains both well-formed classical and ecclesiastical Latin pronunciations with stress and syllable markers, but only for the dictionary forms *contabulō* /konˈta.bu.lo:/ and *dēmonstrō* /de:ˈmon.stro:/.

We must emphasize that we are not condemning the use of machine-generated pronunciations. For many languages, e.g. Spanish and Latin, the spelling of a word reflects its pronunciation, so generated pronunciations are likely to be accurate. Indeed, the existence of pronunciation templates such as `{{fr-IPA}}`, mentioned above, are well-researched additions to Wiktionary that alleviate the need for humans to manually input IPA pronunciations, thus reducing the potential for human error. We fully support the use of these templates (though they make our parsing job harder), and we would love to see them standardized across all Wiktionary editions, so that editions such as the Malagasy edition can benefit from contributions to the English edition (or any other edition, for that matter).

We do caution researchers that the data contained in crowd-sourced resources such as Wiktionary may not be thoroughly vetted for accuracy, as we have discovered. Fortunately, the openness of these crowdsourced data allows for community members to quickly intervene when problematic data is found. One especially poignant example in recent news is the Scots Wikipedia, a large portion of which was recently revealed to be written by an American teenager who is not a Scots speaker.[6] Essentially, this teenager translated English articles into "Scots" by systematically rewriting English words to sound as if they were spoken with a Scottish accent, in the same vein as some of the Latin "IPA" pronunciations in the Malagasy Wiktionary.

4 Visualizing Syllabification

IPA has the ability to mark syllable boundaries (.) as well as primary (ˈ) and secondary (ˌ) stress. Words in some languages, e.g. Malay, do not have stress, and sometimes stress can be double marked (ˈˈ) for extra stress. We first quantify IPA stress and syllabification in our extracted dataset then present multilingual experiments on predicting syllabification and stress.

We develop a visualization technique to understand the distribution of words in each language that contain syllable boundaries (Figure 3). These bubble charts plot the number of characters in a word (x-axis), the percentage of words containing syllable markers (y-axis), and the number of words

[4] https://en.wikipedia.org/wiki/List_of_Wiktionaries

[5] Inflected words have their own Wiktionary entry, which can exponentially increase the number of pronunciations.

[6] https://www.reddit.com/r/Scotland/comments/ig9jia/ive_discovered_that_almost_every_single_article

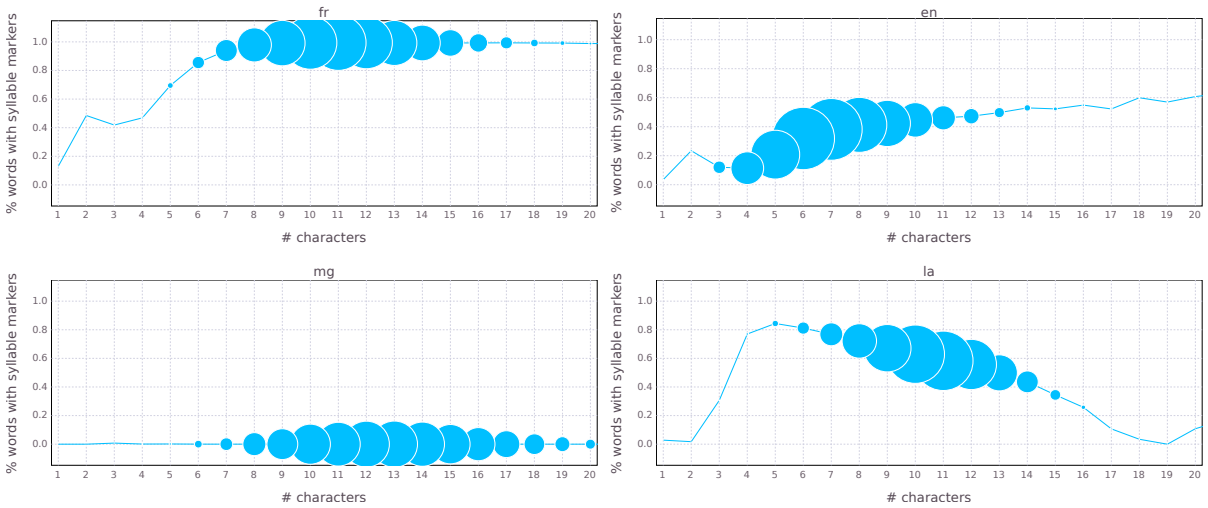

Figure 3: Percentage of French, English, Malagasy, and Latin words containing syllable markers, by length of word. The size of the points indicates the number of words and cannot be compared among graphs.

in these categories (size of the dot). These charts can help researchers to quickly quantify the presence of syllable markers, one component of high-quality IPA pronunciations. We consider a word to be syllabified if it contains any of the following symbols: . ˈ ˌ

Ideally, one should see that the longer the word, the higher the percentage of words that have syllables marked. French is a perfect example of this: once words reach 9–10 characters in length, they all contain syllable markers. By examining these plots, we can easily identify examples of problematic IPA syllabification in Malagasy (`mg`) and Latin (`la`) words. For Malagasy words, syllable boundaries simply do not exist. For Latin words, we see an unusual negative-slope curve, where words around 4–6 characters in length are more likely to have syllables marked, but longer words are less likely to have syllable boundaries marked. This analysis actually is consistent with our earlier finding in Section 2: because Latin is a highly inflected language, the dictionary forms contain high-quality IPA, but the overwhelming number of pronunciations are actually machine-generated for inflected forms, which may not have the syllables marked. English is a middle ground in terms of quality. While we see the expected upward slope as the length of the word increases, the percentage of words with syllable markers never approaches 100%. A manual review of several English pronunciations indicates that annotators simply did not include syllable boundaries for many English words. Further analyses could shed light on the reasons for the negligence of the annotators, or other phenomena that might explain the lack of syllable markers.

5 Experiments

In this section, we present experiments on multilingual syllable and stress prediction. In the linguistics literature, many studies have shown that awareness of syllable boundaries can improve word recognition performance in children (e.g. McBride-Chang et al., 2004; Plaza and Cohen, 2007; Güldenoğlu, 2017). Speech syllabification is also a common step in a speech recognition pipeline. Syllabification of text is not a new task, and has been explored via a variety of methods, including rule-based and grammar-based approaches (e.g. Weerasinghe et al., 2005; Müller, 2006) and data-driven approaches (e.g. Bartlett et al., 2008; Nicolai et al., 2016; Gyanendro Singh et al., 2016). However, previous work has focused primarily on a handful of languages, and some focus on orthographic syllabification rather than phonemic segmentation. Some use CELEX (Baayen et al., 1996), a popular dataset containing syllabified text, but it only contains syllabified words in English, German, and Dutch. In contrast, our extracted pronunciation lexicon is a unique multilingual resource that allows for developing and evaluating models and approaches on the new combined task of massively multilingual IPA syllabification *and* stress prediction across hundreds of languages. In this task, given unmarked IPA, a model must insert syllable markers or stress markers at the appropriate locations.

Data For our task, we filter our pronunciation dataset to keep only IPA containing syllable boundaries or stress markers,[7] so that we have ground truth for our model. This resulted in 93,206 IPA pronunciations across 174 languages, which we split into a 80-10-10 train-dev-test stratified split (same proportion of languages in each set).

Models We first build a baseline: a multilingual character BiLSTM sequence tagger with 256 hidden size (B) that predicts both stress and syllabification (Str & Syl) or syllabification alone (Syl). The data is preprocessed such that each IPA character is labelled with 0 for no stress or syllable, 1 for primary stress (ˈ), 2 for secondary stress (ˌ), and 3 for syllable boundary (.). We include a language token so the model will incorporate knowledge of the language. For example:

IPA:	/ˌɪn.fluˈɛn.zə/
Input:	`eng ɪ n f l u ɛ n z ə`
Output:	`0 2 0 3 0 0 1 0 3 0`

For comparison, we experiment with two modern seq2seq models: the default encoder-decoder model (S) in OpenNMT-py (Klein et al., 2017), and the same model with copy attention (SC) (See et al., 2017). In this scenario, we formulate syllabification and stress prediction as a sequence generation task, where the input is an unstressed, unsyllabified IPA, and the output is the original IPA sequence containing both stress and syllable markers.

We then treat syllabification and stress prediction in a pipelined approach (Syl → Str), where the first model (B or SC) will predict syllable boundaries, and then a second model will predict the stress. Stress classification is a 3-class classification problem: given a syllable, predict primary stress, secondary stress, or no stress. The structure of this stress classifier is also a BiLSTM, where the hidden state of the syllable in question is passed to a dense feed-forward layer, then a softmax.

5.1 Results

A summary of experimental results is in Table 1. The baseline BiLSTM model performs consistently worse than the seq2seq models. This is somewhat surprising, since the seq2seq task is a more challenging task: the model must generate the IPA characters along with stress and syllable markers. However, the seq2seq model is able to generate the correct sequence of IPA characters, minus stress and syllable markers, in 95% (for regular attention) and 99% (for copy attention) of test examples, alleviating our concerns and proving the effectiveness of copy attention for this task.

Model	Acc	CED	5Acc	5CED
B Syl	68	.48	—	—
SC Syl	79	.42	96	.11
B Syl → Str	53	.88	—	—
SC Syl → Str	31	1.13	—	—
B Str & Syl	52	.89	—	—
-Str	68	.49	—	—
S Str & Syl	69	.72	89	.25
-Str	77	.47	93	.16
SC Str & Syl	74	.54	92	.17
-Str	81	.35	95	.11

Table 1: Results on the syllabification and stress prediction tasks. See Section 5 for abbreviations. Acc is 1-best accuracy, 5Best is 5-best accuracy (is the gold in the top 5 hypotheses?), CED is mean character edit distance.

The pipeline approach performs substantially worse than the multi-task approach. In the pipeline, the syllabification model first predicts the syllable boundaries, then the stress classifier produces a classification for each syllable. We find that with the pipeline approach, it is impossible to improve upon the first step in the pipeline. Thus, if the syllabification step does not correctly identify syllable boundaries, the final pronunciation will never be correct, even if the stress is correctly predicted for each syllable.

Finally, multi-task training on both syllabification and stress marking improves performance over syllabification alone. We believe this is because stress and syllable prediction are two somewhat overlapping tasks. If a model can label stress, then it should have some notion of where syllables are. The (-Str) rows in Table 1 show performance on syllabification by evaluating the output of the multi-task model preprocessed to replace all stress marks with syllable boundaries.

The large majority of languages in our dataset can be considered low-resource, a specific interest of our experiments. 154 of the 174 languages have much fewer than 466 training examples (0.5% of the entire dataset), yet the average accuracy on these languages is an impressive 67% for syllabifi-

[7] A stress marker can server as a syllable boundary, e.g. for the English word *consume* /kənˈsum/.

cation (B Str & Syl - Str) and 51% for both syllabification and stress prediction (B Str & Syl). This highlights the contribution of other languages in a single massively multilingual model trained to do both tasks. Other researchers have found that good performance on syllabification requires much more data than this (Nicolai et al., 2016). We highlight the fact that many of the languages have less than 10 test examples and can be considered truly low-resource; the contribution of many other languages allows our multilingual models to predict the correct pronunciation with minimal training data in a specific language. Though we find that multilingual training helps for low-resource languages, it can also help with high-resource languages: in the SC Str & Syl scenario, a model trained only on French obtained 92.1% on the French test words, compared to the multilingual model at 98.1% accuracy. Full tables of results, along with code to reproduce our experiments, is available at https://github.com/wswu/syllabification.

6 Conclusion

We extracted the largest dataset of IPA pronunciations to date, by combining IPA from the French, Spanish, Malagasy, Italian, and Greek editions of Wiktionary along with existing pronunciations from the English edition, totaling to 5.3 million pronunciations. We developed a visualization method for examining syllabification in large datasets, which can give indications about the quality of IPA pronunciations. We experiment on the new combined task of massively multilingual prediction of syllabification and stress using a variety of models and approaches, showing success with a multitask multilingual sequence-to-sequence model. We hope our dataset and analysis methods will be useful for researchers in a variety of disciplines.

We envision our newly extracted pronunciation dataset to be especially useful for researchers interested in lexicography and spoken language technologies. In terms of lexicography, this dataset is a unique comparable corpus containing annotations from several editions of Wiktionary, each representing a distinct population of speakers. In several cases, the same pronunciation is supplied by multiple editions, and some editions use phonetic rather than phonemic IPA. Future work can address questions such as: When and why might different editions disagree on a pronunciation? Why do some words have pronunciations and others don't? In addition, we would like to investigate the use of our pronunciation dataset in language learning of core vocabulary of low-resource languages (Wu et al., 2020) and modeling etymology relationships between words (Wu et al., 2021).

References

R. H. Baayen, R. Piepenbrock, and L. Gulikers. 1996. The celex lexical database (cd-rom).

Susan Bartlett, Grzegorz Kondrak, and Colin Cherry. 2008. Automatic syllabification with structured SVMs for letter-to-phoneme conversion. In *Proceedings of ACL-08: HLT*, pages 568–576, Columbus, Ohio. Association for Computational Linguistics.

Kyle Gorman, Lucas F.E. Ashby, Aaron Goyzueta, Arya McCarthy, Shijie Wu, and Daniel You. 2020. The 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 40–50, Online. Association for Computational Linguistics.

Birkan Güldenoğlu. 2017. The effects of syllable-awareness skills on the word-reading performances of students reading in a transparent orthography. *International Electronic Journal of Elementary Education*, 8(3):425–442.

Loitongbam Gyanendro Singh, Lenin Laitonjam, and Sanasam Ranbir Singh. 2016. Automatic syllabification for Manipuri language. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 349–357, Osaka, Japan. The COLING 2016 Organizing Committee.

G. Klein, Yoon Kim, Y. Deng, Jean Senellart, and Alexander M. Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *ArXiv*, abs/1701.02810.

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 12th Language Resources and Evaluation Conference*, pages 4223–4228, Marseille, France. European Language Resources Association.

Catherine McBride-Chang, Ellen Bialystok, Karen KY Chong, and Yanping Li. 2004. Levels of phonological awareness in three cultures. *Journal of Experimental Child Psychology*, 89(2):93–111.

Karin Müller. 2006. Improving syllabification models with phonotactic knowledge. In *Proceedings of the Eighth Meeting of the ACL Special Interest Group on*

Computational Phonology and Morphology at HLT-NAACL 2006, pages 11–20, New York City, USA. Association for Computational Linguistics.

Garrett Nicolai, Lei Yao, and Grzegorz Kondrak. 2016. Morphological segmentation can improve syllabification. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 99–103, Berlin, Germany. Association for Computational Linguistics.

Monique Plaza and Henri Cohen. 2007. The contribution of phonological awareness and visual attention in early reading and spelling. *Dyslexia*, 13(1):67–76.

Franck Sajous, Basilio Calderone, and Nabil Hathout. 2020. ENGLAWI: From human- to machine-readable Wiktionary. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3016–3026, Marseille, France. European Language Resources Association.

Tim Schlippe, Sebastian Ochs, and Tanja Schultz. 2010. Wiktionary as a source for automatic pronunciation extraction. In *INTERSPEECH*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Ruvan Weerasinghe, Asanka Wasala, and Kumudu Gamage. 2005. A rule based syllabification algorithm for Sinhala. In *Second International Joint Conference on Natural Language Processing: Full Papers*.

Winston Wu, Kevin Duh, and David Yarowsky. 2021. Sequence models for computational etymology of borrowings. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4032–4037, Online. Association for Computational Linguistics.

Winston Wu, Garrett Nicolai, and David Yarowsky. 2020. Multilingual dictionary based construction of core vocabulary. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4211–4217, Marseille, France. European Language Resources Association.

Winston Wu and David Yarowsky. 2020a. Computational etymology and word emergence. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3252–3259, Marseille, France. European Language Resources Association.

Winston Wu and David Yarowsky. 2020b. Wiktionary normalization of translations and morphological information. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4683–4692, Barcelona, Spain (Online). International Committee on Computational Linguistics.