Syllable Structures Across Arabic Varieties

Abdelrahim Qaddoumi[1,2] **Jordan Kodner**[1,2] **Salam Khalifa**[1,2,3]
Ellen Broselow[1] **Owen Rambow**[1,2]
[1]Institute for Advanced Computational Science; [2]Department of Linguistics
Stony Brook University
[3]Computational Approaches to Modeling Language (CAMeL) Lab, NYU Abu Dhabi
{first.last}@stonybrook.edu

Abstract

This study compares the syllable structures of nine Arabic varieties from Wiktionary, using a computational syllabifier. It further investigates methods for learning syllable boundaries in unsyllabified words transcribed in the International Phonetic Alphabet (IPA). The syllabification algorithm is evaluated under three conditions: (i) **Default**, employing fixed rules; (ii) **Joint**, learning onsets and codas across all varieties collectively; and (iii) **Per-variety**, learning onsets and codas specific to each variety. Results indicate that the default configuration yields the highest accuracy, ranging from 97.05% to 100%. The per-variety approach achieves 90.64% to 100% accuracy, while the joint approach ranges from 84.63% to 94.74%. A cross-variety analysis using Jensen-Shannon divergence reveals three principal groupings: Egyptian, Hejazi, and Modern Standard Arabic are closely related; Levantine and Gulf varieties constitute a second cluster; and Juba Arabic, Maltese, and Moroccan emerge as outliers. A cleaned dataset encompassing all nine varieties is also provided.

1 Introduction

One of the fundamental elements of phonology is the syllable (Goldsmith, 2011), abstract constituents of the mental representation of sound structure (Al-Ani and May, 1973). While they are not always detectable from the audio signal, they have been invaluable for the study of phonological processes such as word-stress assignment (Broselow, 2017).

Syllables play a prominent role in early language acquisition. Infants as young as four days old perceive syllables and can discriminate words based on syllable length (Jusczyk and Derrah, 1987; Bijeljac-Babic et al., 1993). Syllables for the basic units for word segmentation, where errors overwhelmingly line up with syllable boundaries (Peters, 1983). Given this, syllables are commonly assumed as the basic unit in the developmental literature on word segmentation, both experimental Saffran et al. (1996) and computational (Lignos, 2011; Fourtassi et al., 2013). Syllable-based models for word segmentation have been shown to outperform phoneme-based models (Schrimpf and Jarosz, 2014).

Word	da.ras.ha	#	dar.ras	#	drUs	#	dars
Structure	Cv.CvC.Cv	#	CvC.CvC	#	CCVC	#	CvCC
Onset	d r h		d r		dr		d
Nucleus	a a a		a a		U		a
Coda	∅ s ∅		r s		s		rs

Table 1: Syllabified examples, in Levantine Arabic, decomposed into onset, nucleus, and coda, with markers for syllable boundary ., and word boundary #. The syllable *structure* is represented with C for consonants, v for short vowels, and V for long vowels. The glosses for the words are <دَرَسها> 'he studied it [f.sg]', <دَرَّس> 'he taught', <دروس> 'lessons', and <دَرس> 'lesson', respectively.

Syllables have also played a similarly important role in word segmentation as an NLP task, forming the basic unit of segmentation in languages with scripts that lack explicit word spacing (Htay and Murthy, 2008; Chormai et al., 2020). In downstream applications, syllable boundaries have been shown to improve the performance of morphological analyzers (Khalifa et al., 2025) and to improve the generalization ability of a Chinese LLM in the presence of phonological ambiguities by introducing subtasks, such as converting orthography to a syllabified form written in IPA and handling tone. In addition, they use LLMs to generate synthetic data containing syllable and tone information. They show that using these subtasks and synthetic data for training their model yields performance gains (Ma et al., 2025).

However, what counts as an acceptable syllable is not identical across all languages, or even di-

Proceedings of the Thirteenth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 250–260
March 29, 2026.

alects of the same language. In this paper, we study syllable structure across Arabic varieties, which diverge considerably in what they accept as a syllable. For example, in Egyptian, it is rare to see syllable structures such as CCvC in loanwords as in <كمسيون> [kum.sjon] 'commission'. Moroccan has triple consonant CCC onset clusters, for example, <استعاد> [stʕa:d] to call for 'help'. Maltese has single vowels as independent syllables, for example, <emozzjona> [ɛ.mɔt.tsjɔ:.na] 'to move'.

Previous studies have examined empirical data on Arabic syllable structure. Hamdi et al. (2005) analyzed the syllable structures of Moroccan, Tunisian, and Lebanese varieties, showing that the frequency of syllable types differs across varieties. For example, Lebanese tends to favor syllables with long vowels, such as CV and CVC, while Moroccan prefers syllable structures with more consonants, like CCvC and CCvCC. Our analysis will include additional varieties and quantify their similarities based on syllable distributions.

Unsurprisingly given the importance of syllables for linguists and developmental researchers, There is a long history of research on automatic syllabification, though this has primarily focused on English. Marchand et al. (2007) compares different algorithms, including Fisher (1996)'s rule-based approach based on Kahn's procedure (Kahn, 2015), as well as several data-driven methods: syllabification by analogy, a look-up procedure, and exemplar generalization. Their results show that data-driven approaches outperform rule-based ones. Wu and Yarowsky (2021) explores multilingual syllabification and stress prediction, using phonemic representations in IPA transcriptions.

In this paper, we automatically extract syllables from a range of Arabic varieties using Kodner (2016)'s SimpleSyllabify, aa simple, language-independent syllabifier (see Section 3) that can run on both orthographic and phonemic representations. This paper makes the following contributions to the study of Arabic syllable structure:

1. We introduce a new IPA dataset derived from Wiktionary as a "silver standard" for Arabic syllabification (Section 4).
2. We use this dataset to evaluate the performance of SimpleSyllabify (Section 5).
3. We use it to perform a cross-variety comparison of syllable inventories (Section 6). We use targeted error analysis (Section 7) to separate algorithmic errors from data artifacts.

2 Linguistic Background

This section provides the linguistic background needed to follow the paper. Our work does not make claims about which linguistic theory is correct; we simply state the theoretical assumption used in this analysis. For terminological convenience, we use the terms *variety* and *varieties* to mean all of the following: Arabic dialects, Modern Standard Arabic (MSA), Juba, and Maltese.

In the following discussion, forms in square brackets ([ ]) represent transcriptions of surface forms (SF) using the International Phonetic Alphabet (IPA), while those between slashes (/ /) represent underlying representations (UR). Arbitrary consonants are represented with C, and short and long vowels with v and V, respectively. Specific short and long vowels are also represented in lower and upper case, for example, short [a] and long [A]. This is sufficient granularity for the present study of Arabic. Breaks between syllables are indicated with a period (.), and word boundaries are indicated by '#'. For example, in South Levantine, <درّس> /darras/ 'he taught' has a geminate: [dar.ras] (CvC.CvC), while <درس> /dars/ 'lesson' has a complex cluster: [dars] (CvCC).

2.1 Syllables

In this paper, we divide syllables into three constituents, the onset, the nucleus, and the coda (Hayes, 2009). The nucleus is the core of the syllable, and usually takes the form of a vowel or a vocalic consonant. The onset is the segment preceding the nucleus, and the coda is the segment following the nucleus. Table 1 demonstrates examples of this definition, it showcases different shapes of syllables in different contexts.

2.2 Arabic Varieties

Arabic is often treated as a single language. However, it is actually a collection of regional varieties that differ in many linguistic dimensions, including their syllable inventories and phonological patterns. These regional varieties are the native varieties of Arabic speakers across the Arabic-speaking world and coexist alongside MSA. MSA is considered the "prestige" variety, as it is used in formal settings, while the regional spoken variety is used day to day. This special phenomenon of a single community using two dialects or varieties is described in the literature as diglossia (Ferguson, 1959).

In this paper, we analyze the following varieties: Egyptian (EGY), Gulf (GUL), Hejazi (HEJ), Juba (JUB), Maltese (MAL), Moroccan (MOR), North Levantine (NLE), South Levantine (SLE), and MSA. MSA is the standard written form across the Arabic-speaking world. The Egyptian Arabic variety is notable for its prominence in media across the Arab world, its wide intelligibility, and it is spoken by 119 million people. Moroccan Arabic, a Maghrebi variety, is spoken by 40 million people and is influenced by contact with Tamazight. Hejazi Arabic, spoken in western Saudi Arabia, and Gulf Arabic, spoken in the states bordering the Persian Gulf, are Arabian Peninsula varieties, each with about 11 million speakers. Levantine encompasses about 60 million speakers and is divided between South Levantine (Jordanian/Palestinian) and North Levantine (Lebanese/Syrian) (Eberhard et al., 2025).

The remaining two varieties are less frequently featured in comparative studies. Located in the peripheries of the Arabic-speaking world, they have been subject to extensive contact with non-Semitic languages. Juba Arabic is an Arabic-based creole spoken by 250,000 people in South Sudan. Maltese, which around 500,000 people speak in Malta (Rosner and Borg, 2022), is an official EU working language descended from medieval Siculo-Arabic. It is a relative of modern North African varieties but with substantial influence from Romance languages (Brincat, 2005). It is notable as one of the few Semitic languages to be written in a Latin script. Taken together, its orthography and institutional status make it atypical relative to Arabic regular diglossia.

2.3 Syllable Structure Variation

Due to diglossia between local varieties and MSA, data for any local variety is relatively scarce. Diglossia presents a research challenge for both NLP and linguistics. In NLP, Arabic regional varieties are low-resource, making it challenging to work with limited data. However, their syllable structures vary in predictable ways, providing ground for cross-variety transfer techniques. From a linguistic perspective, this is an ideal setting for quantitatively comparing features of linguistic variety, such as syllable structures.

Since Arabic varieties are closely related yet differ in the syllable structures that they employ, they constitute a useful comparison set in the phonological study of syllables. We illustrate some of the differences here. All Arabic varieties in this paper, allow at least the following syllable structures: CV, and CVC, except for Juba Arabic, which lacks long vowels. For a complete literature summary, see Table 4.

JUB additionally permits CCv (Manfredi and Petrollino, 2013). Onsets are mandatory in Arabic, so underlyingly vowel-initial syllables are realized with a glottal stop [ʔ] (Broselow, 2017). This is generally true in MAL as well, making V marginal as a phonotactic template (Galea and Ussishkin, 2018).

While CC and CCC are possible onsets in MOR in analyses such as (Kiparsky, 2003), there is still debate about how to analyze syllables in MOR. For example, Shaw et al. (2009) use articulatory measures to argue for syllables consisting only of simplex-onsets. Thus, /skru/ 'his ploughshares' surfaces as [s.kru] "C.CCv" instead of [skru] "CCCv". MOR has been shown to have complex-onset varieties such as CCv/CCvC, CCCv, and CCCvC in the surface form (Hamdi et al., 2005).

SLE and NLE allow for CCvC and CCVC structures (Hamdi et al., 2005). Moreover, GLF allows complex-onset CCV. Standard Maltese permits syllables of the form (C)(C)(C)V(V)(C)(C), so we find a range of complex onsets and codas in the language (Galea and Ussishkin, 2018).

Finally, Arabic varieties impose various restrictions on syllable types depending on word position. Some Levantine varieties allow CC onsets only in word-initial position; EGY allows CVC, CvCC only in word-final position (Broselow, 2017). Despite not using syllable position in the syllabification, the high accuracy in our data (Table 3) suggests that position-conditioned syllables are either rare in the data or are handled correctly by the default setting.

Table 4 illustrates how syllable-type inventories can differ across varieties. Examples of the differences are: the availability of complex onsets (e.g., CCvC/CCVC), the distribution of superheavy syllables in word-final position, permissible onsets, and edge-sensitive restrictions on heavy codas. These are precisely the values that matter for the syllabification algorithm. Examples of these cases are explained in further detail in Section 3

3 Syllabification Method

To obtain the actual syllabification, we used the SimpleSyllabify syllabification script by Kod-

ner (2016)[1] which identifies syllable nuclei and then inserts syllable boundaries according to the Maximum Onset Principle (Kahn, 2015), a widely understood generalization from phonology (Goldsmith, 2011). The SimpleSyllabify is intended to run on orthographic representations, but, as there are no standard orthographies for Arabic varieties, standard Arabic orthography usually omits short vowels, and IPA has wide acceptance among linguists, we use IPA-like transcriptions from Wiktionary, as described in Section 4.1.

The script requires a list of characters that represent syllable nuclei (usually vowels, but possibly characters representing syllabic sonorants).[2] Given this set of characters, the script proceeds by identifying all of the syllable nuclei in a word. It then considers each consonant cluster to the left of a nucleus as a potential onset. Maximum Onset Principle states that the largest string of onset consonants in this cluster that can licitly be incorporated into the onset should be, and the remainder is incorporated into the previous syllable's coda. All word-initial consonants are added to the initial syllable's onset, and all word-final consonants are added to the final syllable's coda.

Since what counts as a licit onset differs from one variety to another, the script defaults to splitting consonant clusters in half, assigning half (plus one in the case of odd-sized clusters) into the onset, with the remainder added to the previous syllable's coda. For example, vCCCCV is syllabified as vCC.CCV and vCCCV are syllabified as vC.CCV.

Alternatively, the script can learn a language-specific set of licit onsets from word-initial onsets in the dataset. However, this may hurt performance if the training set is small, leading to underlearning, or if the variety allows word-initial onsets that are banned word-medially. Unfortunately, these are both problems in our case, and learning from word-initial onsets dramatically lowers performance (See Table 3). A third option is for the user to provide an explicit list of valid onsets. However, specifying variety-specific onsets up-front undermines the spirit of our experiments. Since syllabification performance is over 96% for all varieties and above 99% for the majority, we find that the default splitting approach is sufficient for our purposes.

It is theoretically possible to learn per-variety syllabification algorithms that account for variety-specific inventories of licit onsets and codas, the sonority of different consonants, and whether the variety allows rising or lowering sonority in onsets or codas. In practice, this is difficult to derive from Wiktionary data because it is limited and noisy. This is why we use a general deterministic algorithm.

4 Data

4.1 Wiktionary

Wiktionary[3] is a free, collaborative online lexicon that provides structured entries for many languages, often including IPA transcriptions, audio, and glosses (Meyer and Gurevych, 2012). It has data for 22 Arabic varieties. From these, we selected the varieties for which there were at least 100 entries with syllabified IPA transcriptions: Egyptian (EGY, 584 words), Gulf (GUL, 636 words), Hejazi (HEJ, 2,919 words), Juba (JUB, 189 words), Maltese (MAL, 17,609 words), Moroccan (MOR, 1,955 words), Modern Standard Arabic (MSA, 16,718 words), North Levantine (NLE, 575 words), and South Levantine (SLE, 5,059 words).[4] We extracted our datasets from a Wiktionary dump (Ylönen, 2022) and processed them into per-variety word lists.

4.1.1 Preprocessing

While Wiktionary is overall a reliable source, it is crowdsourced, and thus the data may be noisy (Sakunkoo and Sakunkoo, 2025). We noticed that IPA conventions vary across contributors, which we address with preprocessing. We remove duplicates and any entries with transcriptions that contain only one or two characters, since they are trivial to syllabify. We further standardize transcriptions as follows: Long vowels are capitalized in order to represent them as a single character, for example, <كاتب> [ka:.tib] 'writer', will be [kA.tib]. Geminates are represented by repeating the same character, for example, <كتّب> [kat:ab] 'he made x write', will be represented as [kat.tab]. Finally, we represent the sibilant affricate segments "ts", "dʒ", and "tʃ" as single consonants "T", "D", "S"

[1] https://github.com/jkodner05/syllabify

[2] The following set of symbols were treated as vowels for the purposes of identifying nuclei: ʊ, a, ɜ, u, e, æ, o, ɑ, i, ɪ, ə, ɔ, ɐ, ɘ, ä, ɛ, ʉ, õ, œ, ɨ, ɒ, ʌ, Ʊ, A, U, E, Æ, O, Ɑ, I, Ɪ, Ə, Ɔ, Ɐ, Ä, Ɛ, Ʉ, Õ, Œ, Ɨ, Ʌ. Lowercase symbols are the IPA vowels used in Wiktionary. We used the upper case symbols for our normalization of long vowels.

[3] https://www.wiktionary.org/

[4] We omit Algerian, Andalusian, Baharna, Chadian, Cypriot, Hassaniya, Iraqi, Libyan, North Mesopotamian, Omani, Sudanese, Tunisian, and Yemeni.

respectively. This decision is supported by the fact that affricates are contour segments, which means that they have two phonetic qualities, but can be treated as a single segment phonologically (Hayes, 2009).

When an entry provides multiple IPA transcriptions, we retain each distinct transcription as a separate surface form for that variety, for example, <شُوكُولَاتَة> [ʃu.kU.lA.ta] or [ʃu.ku.lA.ta] 'chocolate'. For the cases where one segment is presented as optional in a single transcription, we always include the optional segment, as in the following example, where <اسمر> /(ɪ)smar/ 'dark' is treated as [ɪs.mar].

Stress marks are removed. All superscripts such as 'ˤ' and subscripts such as '̪' are removed as well, since these correspond to annotations or to phonological distinctions which do not affect Arabic syllabification (Association, 1999), for example, the tie bar '͡' and pharyngeal diacritics 'ˤ' in <آية الله> /ʔA.ja.tu͡lˤ.lˤAh/ 'sign of God'. We remove [ ] and / /, since they do not represent segments.

4.2 "Bronze", "Silver", and "Gold" Standard

We compare five tiers of postprocessing from the raw Wiktionary data to quantify (i) how well the rule-based syllabifier matches crowdsourced data and (ii) how much of the observed mismatch is due to data labeling errors instead of syllabifier errors. **Wiktionary** is raw data with the website's syllable annotations preserved as-is. **Predicted** is the output of the syllabifier applied to Wiktionary IPA with original syllabification information discarded. **Bronze** excludes the three easy-to-identify mislabeling errors. The data excluded is described in the following subsection and reported in Table 2.

Silver retains **Bronze** entries and automatically corrects mislabeling as described below. **Gold** retains entries where Predicted matches Wiktionary, and manually corrects mislabeling by a native speaker.

4.2.1 Bronze Data

To generate the Bronze data, we remove the following labeling errors in the Wiktionary data. The first is Under-Segmentation (US) errors, where transcriptions clearly have an insufficient number of syllable boundaries N_{syll} compared to number of vowels N_{vowel}. Since there should be a boundary somewhere between every pair of vowels, we can exclude words where $(N_{\text{syll}} < N_{\text{vowel}} - 1)$ (Hayes, 2009). An annotation like [ka.taba] can be excluded because it has a structure CvCvCv with three vowels, but only one boundary.

The second type is Geminate (G) errors, which occur when transcriptions contain geminate sounds in onsets rather than splitting them into a preceding coda and a following onset, as they should be treated in Arabic (Farwaneh, 2009). For example, [ka.ttab] can be excluded, since it should be annotated as [kat.tab]. The third type is the Single Consonant (C) errors, which occur when syllables are annotated without a nucleus (v or V). For example, [k.at.tab] and [ka.t.tab] are excluded on these grounds.

4.2.2 Silver Data

To generate the Silver data, we first include any words where the predicted syllables match the reference. For items with mismatches, we apply targeted heuristics to recover reliable, predictable errors. For the single-consonant syllable error, we reattach the single-C syllable in the reference. If it is word-initial, it is attached to the subsequent syllable [k.at.tab] → [kat.tab]; if it is word medial, it will be attached to the previous syllable [ka.t.tab] → [kat.tab].

Geminate errors are corrected by splitting the geminate and attaching it to the previous syllable if it is word-medial [ka.ttab] → [kat.tab]. For under-segmentation errors, we look for words that do not contain triple-consonant sequences (CCC) in their structure, as these are generally more prone to error. If the word does not contain CCC, then we apply our syllabification to it.

4.3 Gold

The author-annotators checked a random sample of 100 words in three varieties (SLE, NLE, EGY). In these samples, we found almost no issues: 0 for EGY, 1 for NLE, and 1 for SLE. We could not manually correct additional varieties because we lacked native speakers. We use this **Gold** tier to check that the **Silver** automatically fixed data matches the manually labeled data. The number of mismatched data points between Gold and Silver is 2 for EGY, 5 for NLE, and 50 for SLE.

5 Evaluation

The evaluation is conducted on the bronze, silver, and gold datasets. These datasets are described in Section 4.2. Data points listed in Table 2 are excluded from the evaluation of bronze only.

Variety	US	US+G	G	C	US+C	G+C
EGY	186	32	3	2	2	0
GUL	184	39	13	0	0	0
HEJ	296	25	0	22	0	0
JUB	74	1	0	1	0	0
MAL	177	9	1	1,288	11	3
MOR	58	2	0	2	0	0
MSA	112	7	0	3	0	0
NLE	106	16	0	11	0	1
SLE	97	13	1	1	0	0

Table 2: Counts of excluded words from bronze by error type. Some words exhibited two error types simultaneously. *US* under-segmentation error. *G* geminate error. *C* single consonant error.

We provide two evaluation metrics: accuracy, computed as the percentage of words whose predicted syllabification matches the annotated reference syllabification exactly, and average edit distance between the predicted and reference syllabifications. For example, if the reference word is <kitten> [kɪ.tən] and the predicted is [kɪtən], then the edit distance is one (insertion of the syllable boundary marker); if the output is [kɪt.ən], the edit distance is 2 (deletion of the incorrect boundary and insertion of the correct one).

From the results summarized in Table 3, we see that across all tiers, the Default setting achieves almost perfect accuracy (97–100%) for every variety, while Joint achieves the lowest accuracy (84–97%). The Per-variety setting is usually similar to Default, but it underperforms for MAL and MOR. Since the Bronze and Silver tiers are automatically derived, performance on these sets could potentially be inflated in principle, so they should be compared with Gold, where the same ranking (Default/Per-variety much greater than Joint) remains.

6 Analysis

We examine the distribution of syllable structures across the different varieties. The goal is to see how syllable structure differs among them. To run the analysis, we generate syllable-shape distributions for five tiers: Predicted (syllabifier output), Wiktionary (raw), Bronze, Silver, and Gold.

We summarize the distribution of the syllables in Table 4. While the distributions captured from the predictions and from the silver annotations are very similar, they are not identical. To better understand the difference between them, we use Jensen-Shannon divergence (JSD), which quantifies the difference between probability distributions, and has been widely used in the analysis of symbolic

Data			Default		Joint		Per-variety	
Tier	**Variety**	**# Words**	**% Acc**	**SED**	**% Acc**	**SED**	**% Acc**	**SED**
B	**EGY**	359	99.44	2.0	90.81	2.3	99.44	2.0
B	**GUL**	398	100.00	0.0	94.74	2.1	100.00	0.0
B	**HEJ**	2,574	99.65	2.2	91.46	2.0	99.22	2.1
B	**JUB**	104	100.00	0.0	93.27	2.0	98.08	2.0
B	**MAL**	16,119	99.15	2.0	84.63	2.0	90.64	2.0
B	**MOR**	1,893	99.74	2.0	85.63	2.0	93.03	2.0
B	**MSA**	16,595	99.96	1.7	87.87	2.1	99.27	2.0
B	**NLE**	441	97.05	1.9	89.34	2.0	97.05	1.9
B	**SLE**	4,947	98.99	2.0	88.56	2.1	98.24	2.0
S	**EGY**	584	99.66	2.0	94.35	2.3	99.66	2.0
S	**GUL**	636	99.84	2.0	96.54	2.1	99.84	2.0
S	**HEJ**	2,919	99.45	2.1	92.22	2.0	99.08	2.0
S	**JUB**	189	100.00	0.0	96.30	2.0	98.94	2.0
S	**MAL**	17,609	98.98	1.9	85.01	2.1	90.89	2.0
S	**MOR**	1,955	99.64	2.0	85.98	2.0	93.15	2.0
S	**MSA**	16,718	99.73	1.1	87.73	2.1	99.05	1.8
S	**NLE**	575	97.04	1.9	90.78	2.0	97.04	1.9
S	**SLE**	5,059	97.98	1.5	87.78	2.0	97.25	1.6
G	**EGY**	584	100.00	0.0	94.52	2.3	100.00	0.0
G	**NLE**	575	97.74	1.9	91.48	2.0	97.74	1.9
G	**SLE**	5,059	98.10	2.0	87.94	2.1	97.43	2.0

Table 3: Variety-level evaluation across three settings and three tiers (B=Bronze, S=Silver, G=Gold). Default, Joint (pooled onsets + codas), and Per-variety (variety-specific onsets + codas) report word-level exact-match accuracy and mean edit distance (SED). Gold is available only for EGY, NLE, and SLE.

sequences (Grosse et al., 2002). JSD is related to the Kullback-Leibler divergence (KLD), so it shares the same mathematical properties, but it differs from KLD in two essential ways: it is symmetrical and smooth, unlike KLD (Fuglede and Topsoe, 2004).

6.1 Data Quality Validation

The primary pattern to highlight in Table 5 is that the Gold results validate the automatic correction stage for the three manually checked varieties. The Silver distributions for SLE, NLE, and EGY are indistinguishable from their corresponding Gold distributions, whereas the Wiktionary distributions differ significantly from Gold for varieties such as EGY and NLE.

Silver is a proxy for Gold These patterns indicate that a large portion of the mislabeled data in Wiktionary is automatically repairable, and that the Silver tier is a good proxy for the Gold tier. These results motivate the use of Silver for the following cross-variety analyses.

Cross-variety Differences We observe notable differences between the varieties, for example, in their relative proportions of Cv and CvC. MSA and JUB prefer the former, whereas MOR, NLE, and SLE favor the latter. The rest do not seem to have a clear preference between Cv or CvC. We also see that some syllable structures are nearly ab-

Varieties		Cv	CvC	CVC	CV	CvCC	CCvC	CCVC	CCV	vC	CCv	CVCC	v	Other
EGY	L.	✓	✓	F	✓	F	X	X	X	X	X	-	X	0.4
	P.	37.6	36.4	12.8	6.8	5.8	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.4
	B.	30.7	34.3	19.4	4.7	10.1	0.3	0.3	0.0	0.0	0.0	0.0	0.0	0.3
	S.	37.5	36.5	12.7	6.8	5.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7
	G.	37.6	36.4	12.8	6.8	5.8	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.4
GUL	L.	✓	✓	✓	✓	F	I	I	I	X	✓	-	X	0.4
	P.	33.9	34.3	13.6	10.1	3.1	0.7	1.9	0.9	0.0	0.5	0.4	0.3	0.4
	B.	34.6	30.3	16.7	8.4	4.9	0.4	3.0	0.5	0.0	0.3	0.7	0.0	0.1
	S.	33.9	34.2	13.6	10.0	3.1	0.7	1.8	0.9	0.0	0.5	0.4	0.3	0.5
HEJ	L.	✓	✓	✓	✓	F	*	*	*	X	*	F	X	0.8
	P.	34.5	37.9	13.5	10.0	3.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8
	B.	32.9	38.8	14.2	9.6	3.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8
	S.	34.5	37.8	13.5	10.0	3.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.9
JUB	L.	✓	✓	X	X	X	*	X	X	*	F	X	✓	0.2
	P.	61.7	29.1	0.0	0.0	0.5	0.0	0.0	0.0	2.6	1.0	0.0	5.0	0.2
	B.	60.9	35.0	0.0	0.0	0.5	0.0	0.0	0.0	2.3	0.9	0.0	0.0	0.5
	S.	61.7	29.1	0.0	0.0	0.5	0.0	0.0	0.0	2.6	1.0	0.0	5.0	0.2
MAL	L.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	2.2
	P.	30.4	30.7	8.0	6.6	2.7	3.6	1.3	2.1	7.4	3.3	0.0	1.8	2.2
	B.	30.7	31.5	8.2	6.5	3.0	3.6	1.4	1.9	8.1	3.4	0.0	0.0	1.8
	S.	30.5	30.4	8.0	6.6	2.8	3.6	1.3	2.0	7.4	3.2	0.0	1.8	2.4
MOR	L.	✓	✓	*	*	✓	✓	*	*	X	✓	X	X	0.3
	P.	19.9	33.6	15.2	10.2	4.7	5.6	4.2	4.5	0.0	1.3	0.4	0.0	0.3
	B.	19.8	33.5	15.2	10.1	5.0	5.7	4.4	4.3	0.0	1.3	0.4	0.0	0.3
	S.	20.0	33.5	15.3	10.2	4.8	5.5	4.2	4.6	0.0	1.3	0.4	0.0	0.3
MSA	L.	✓	✓	✓	✓	✓	X	X	X	X	-	✓	X	0.5
	P.	38.5	30.7	12.4	13.5	4.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5
	B.	38.6	30.7	12.4	13.5	4.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4
	S.	38.6	30.7	12.4	13.5	4.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5
NLE	L.	✓	✓	✓	✓	✓	*	I	*	X	-	✓	X	0.9
	P.	28.3	33.5	12.6	11.9	4.5	2.5	2.4	2.5	0.0	0.5	0.5	0.0	0.9
	B.	24.1	33.4	15.6	11.8	6.9	2.0	2.7	1.6	0.0	0.0	0.7	0.0	1.2
	S.	28.3	32.5	12.9	12.0	5.1	2.5	2.0	2.4	0.0	0.4	0.5	0.0	1.5
	G.	28.2	32.7	12.9	12.1	5.3	2.4	1.9	2.4	0.0	0.6	0.5	0.0	1.1
SLE	L.	✓	✓	✓	✓	✓	*	*	*	X	*	-	X	0.1
	P.	30.8	45.3	8.7	8.4	2.9	2.1	0.8	0.8	0.0	0.3	0.0	0.0	0.1
	B.	31.1	45.1	8.9	8.5	3.4	1.4	0.7	0.7	0.0	0.0	0.0	0.0	0.2
	S.	31.0	44.9	8.8	8.4	3.4	1.4	0.7	0.7	0.0	0.0	0.0	0.0	0.7
	G.	30.9	45.1	8.8	8.4	3.8	1.4	0.7	0.7	0.0	0.0	0.0	0.0	0.2

Table 4: Distribution of syllable shapes reported in linguistics literature (*L.*), extracted from syllabifier predictions (*P.*) and dataset tiers (*B./S./G.*) for each variety. *Other* sums all remaining rare types of syllable shapes. 'F' means word-final, 'I' means word-initial, '*' marginal could be for certain conditions such as sonority or certain segments or loanwords, X means it is illicit by the literature, ✓means it exists based on the literature, and '-' means the literature does not mention it. Egyptian syllables (McCarthy, 1979; Aquil, 2013), Gulf (Qafisheh, 1977; Al-Qenaie et al., 2011), Hejazi (Alzaidi et al., 2019; Bokhari, 2020), Juba (Miller, 2006), Maltese (Galea and Ussishkin, 2018), for Moroccan (Boudlal, 2001; Benhallam, 1989; Shaw et al., 2009; Heath, 2020), MSA (Halpern, 2009; Ryding, 2005), North Levantine (Kelly, 2021), and South Levantine (Rakhieh, 2009).

sent in some varieties; for example, CCV, and CCv syllables are prominent only in MOR, and MAL, and CVC and CV syllables are absent from JUB. The results are consistent with prior descriptions of these varieties. MAL, MOR, NLE, and SLE show more complex (CC) onsets than other varieties.

Tier Differences For each tier in Table 5, we compute the distribution of syllable shapes. Then we measure pairwise similarity using JSD. Low JSD values mean distributions are more similar.

From Table 5 we see that Wiktionary's distribution can differ substantially from the predicted distribution. This disagreement is variety-dependent, ranging from near-zero (e.g., MSA) to large values (e.g., EGY, JUB, GUL). This range in values suggests that Wiktionary's syllabification is broadly consistent with the syllabifier's predictions for some varieties, whereas others' are not. These differences may be due to annotation practice (e.g., treatment of glides, epenthetic glottal stops, or geminates) or to systematic entry noise.

The other pattern we noticed is that Predicted is very close to Silver (≈0) for all varieties. The result means that the Silver tier is distributionally almost identical to the syllabifier's output, indicating that the automatic fixes are correcting Wiktionary's inconsistencies without changing the syllabifier's overall syllable structure profile, where we see a nontrivial Bronze–Silver difference in EGY, GUL, JUB, MAL, and NLE.

Lang	W-P	W-B	W-S	W-G	P-B	P-S	P-G	B-S	B-G	S-G
EGY	0.096	0.090	0.093	0.096	0.010	0.001	**0.000**	0.012	0.010	0.001
GUL	0.077	0.078	0.076	—	0.006	**0.000**	—	0.006	—	—
HEJ	0.015	0.014	0.013	—	**0.000**	**0.000**	—	**0.000**	—	—
JUB	0.084	0.104	0.084	—	0.021	**0.000**	—	0.021	—	—
MAL	0.001	0.009	0.001	—	0.008	**0.000**	—	0.008	—	—
MOR	0.006	0.006	0.005	—	**0.000**	**0.000**	—	**0.000**	—	—
MSA	**0.000**	**0.000**	**0.000**	—	**0.000**	**0.000**	—	**0.000**	—	—
NLE	0.047	0.039	0.039	0.043	0.006	0.001	0.001	0.005	0.005	**0.000**
SLE	0.005	0.002	**0.000**	0.002	0.002	0.003	0.002	0.001	**0.000**	0.001

Table 5: JSD across dataset tiers (rounded to 3 decimals). P=Predicted (syllabifier output), W=Wiktionary (raw), B=Bronze (matches+exclude known mislabels), S=Silver (matches+automatic fixes), G=Gold (matches+manual fixes; available only for SLE, NLE, EGY). **Metric.** Each entry is the Jensen–Shannon divergence (JSD) between syllable distributions.

	EGY	GUL	HEJ	JUB	MAL	MOR	MSA	NLE	SLE	Mean
EGY	—	0.020	0.004	0.128	0.081	0.074	0.008	0.036	0.017	0.046
GUL	0.020	—	0.017	0.144	0.054	0.034	0.019	0.011	0.013	0.039
HEJ	0.004	0.017	—	0.140	0.081	0.069	0.005	0.032	0.015	0.045
JUB	0.128	0.144	0.140	—	0.126	0.233	0.142	0.185	0.140	0.155
MAL	0.081	0.054	0.081	0.126	—	0.063	0.084	0.054	0.061	0.076
MOR	0.074	0.034	0.069	0.233	0.063	—	0.074	0.015	0.045	0.076
MSA	0.008	0.019	0.005	0.142	0.084	0.074	—	0.033	0.023	0.049
NLE	0.036	0.011	0.032	0.185	0.054	0.015	0.033	—	0.019	0.048
SLE	0.017	0.013	0.015	0.140	0.061	0.045	0.023	0.019	—	0.042

Table 6: Cross-variety distance matrix (Silver): JSD over syllable distributions. Entries are Jensen–Shannon divergences (JSD) between varieties' Silver tier syllable distributions. The **Mean** column reports each variety's mean JSD to the other varieties (excluding self).

6.2 Cross-variety Analysis

Table 6 reports Jensen–Shannon divergence (JSD) values of the syllable distributions of Silver-tier varieties. JSD is a symmetric measure of distributional dissimilarity, with lower values indicating greater similarity between distributions.

Variety Clusters The general conclusion from the data is that the distances are mostly minimal outside of JUB, MAL, and MOR. The results can be split into three clusters. The first one includes HEJ-EGY, HEJ-MSA, and EGY-MSA. The second cluster of varieties includes GUL, NLE, and SLE; these three varieties are closely related. The last cluster, with its most significant divergence, includes JUB, MAL, and MOR; these varieties are the furthest from the others.

Centrality (mean column) The Mean column provides an overview of each variety's centrality. GUL is identified as the most central variety (Mean 0.039), whereas JUB is the most distant from the mean (0.155), followed by MAL and MOR both (0.076). JUB exhibits the lowest similarity to all other varieties. MAL is the second least similar to most varieties, except in the cases of JUB and MOR, where NLE is the second least similar to JUB, and MSA is the second least similar to MOR.

These results align with the linguistic literature, as the following varieties are generally considered distinct from other Arabic varieties. For example, JUB is creole; Italian and other Romance languages heavily influence MAL, and Tamazight heavily influences MOR and has a very unique consonant cluster compared to other varieties.

7 Error Analysis

SimpleSyllabify's errors are primarily of one type, which is the consistent division of word-internal CCC clusters into C.CC. As discussed in Section 3, this is the default behavior of the algorithm when no set of licit onsets is learned or prespecified. Nevertheless, performance is very high, indicating that this is rarely a problem. For example, in NLE <برتقان> [birt.ʔAn] 'oranges' was predicted as [bir.tʔAn] and in Moroccan, <أربعة> [ʔarb.ʕa] 'four' was predicted as [ʔar.bʕa].

To further refine the syllabifications, we could employ a more complex syllabification algorithm grounded in additional phonology that uses sonority hierarchies, in addition to the Maximum Onset Principle, to adjudicate between syllabifications without explicitly specifying how a variety should split consonant clusters (Clements, 1990). However, sonority hierarchies can also vary cross-linguistically, meaning that segments tend to be arranged according to the Sonority Sequencing Principle, generally with rising sonority in onsets and falling sonority in codas. A simple Sonority Scale from most sonorous to least: vowel < glide < liquid < nasal < obstruent (fricative, stop) (Selkirk, 1984; Clements, 1990).

Second type of error is when there is a glide followed by a consonant, such as in <دويبة> [du.wajb.ba] 'a small animal', the syllabifier generates [du.waj.bba]. The third error type occurs when there are diphthongs, which are two different vowels that share the same syllable nucleus; for example, the reference for JUB has <kweys> [kweis] 'good', and the predicted is [kwe.is]. While SimpleSyllabify has the means to capture these syllables, it requires additional language-specific parameters to be specified.

8 Conclusion and Future Work

We study the problem of syllabification of multiple Arabic varieties from their unsyllabified surface forms written in IPA. We use Maltese, Egyptian, South Levantine, North Levantine, Moroccan, Hejazi, Gulf, and Modern Standard Arabic to evaluate the syllabification algorithm on a wide geographic range and the Arabic continuum. We compare the algorithm's results to the syllable structures described in the literature. Our method shows that applying a non-probabilistic rule-based algorithm is highly successful at syllabifying all Arabic varieties. Then we compare results from using a single syllabifier across all varieties versus one per variety, showing that a general syllabification algorithm performs better for these Arabic varieties. Another thing is how practical this syllabification algorithm is for cleaning data by comparing where it diverges. This allows us to understand the syllable structures in these varieties better.

Future work may involve extending this analysis to a broader range of languages to determine whether the results are consistent with those of the current study. Another avenue for exploration is incorporating sonority ranking into the syllabifier to assess potential improvements in performance. Sonority ranking could be learned using a Bottom-Up Factor Inference Algorithm (BUFIA) (Chandlee et al., 2019), a grammar inference algorithm that infers the most general set of forbidden constraints from positive data. Additionally, methods for learning onsets and codas directly from data, without relying on a sonority hierarchy, warrant investigation.

The identification of different Arabic varieties can be investigated based on their syllable structures and distributions. Furthermore, incorporating this information into neural networks may facilitate data generation for low-resource varieties by augmenting Modern Standard Arabic data, thereby enabling transfer learning.

Acknowledgments

We thank Jeffrey Heinz for his helpful discussion and feedback. We also thank the anonymous reviewers for their feedback. Qaddoumi and Rambow gratefully acknowledge support from the Institute for Advanced Computational Science at Stony Brook University.

Limitations

One of the main limitations of this distributional analysis is that we are looking at the frequency of syllable structures across types rather than the actual token frequency. For example, it can be that the types with CV are less used in real life than those with a different syllable structure because many of the Arabic varieties delete certain short vowels in open syllables. Unfortunately, this is not something we can solve for now, as there is no real frequency for tokens for Arabic varieties for this type of data.

References

Salman Al-Ani and David May. 1973. The phonological structure of the syllable. In Salman Al-Ani, editor, *Readings in Arabic Linguistics*, pages 113–125. Indiana University Linguistics Club, Bloomington.

Shamlan Al-Qenaie and 1 others. 2011. *Kuwaiti Arabic: A socio-phonological perspective*. Ph.D. thesis, Durham University.

Muhammad Swaileh Alzaidi, Yi Xu, and Anqi Xu. 2019. Prosodic encoding of focus in hijazi arabic. *Speech Communication*, 106:127–149.

Rajaa Aquil. 2013. Cairne arabic syllable structure though different phonological theories. *Open Journal of Modern Linguistics*, 3(3):259–267.

International Phonetic Association. 1999. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press.

Abderrafi Benhallam. 1989. Moroccan arabic syllable structure. *Revue Langues et Littératures*, 8:177–191.

Ranka Bijeljac-Babic, Josiane Bertoncini, and Jacques Mehler. 1993. How do 4-day-old infants categorize multisyllabic utterances? *Developmental psychology*, 29(4):711.

Hassan Abdulrashed Bokhari. 2020. *A comprehensive analysis of coda clusters in Hijazi Arabic: An optimality-theoretic perspective*. Indiana University.

Abdelaziz Boudlal. 2001. *Constraint interaction in the phonology and morphology of Casablanca Moroccan Arabic*. Ph.D. thesis, Rutgers University.

Joseph M Brincat. 2005. Maltese–an unusual formula. *MED Magazine*, 27.

Ellen Broselow. 2017. Syllable structure in the dialects of arabic. *The Routledge handbook of Arabic linguistics*, pages 32–47.

Jane Chandlee, Remi Eyraud, Jeffrey Heinz, Adam Jardine, and Jonathan Rawski. 2019. Learning with partially ordered representations. In *Proceedings of the 16th Meeting on the Mathematics of Language*, pages 91–101, Toronto, Canada. Association for Computational Linguistics.

Pattarawat Chormai, Ponrawee Prasertsom, Jin Cheevaprawatdomrong, and Attapol Rutherford. 2020. Syllable-based neural Thai word segmentation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4619–4637, Barcelona, Spain (Online). International Committee on Computational Linguistics.

George N Clements. 1990. The role of the sonority cycle in core syllabification. *Papers in laboratory phonology*, 1:283–333.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2025. *Ethnologue: Languages of the World*, 28 edition. SIL International, Dallas, Texas. Online edition.

Samira Farwaneh. 2009. Toward a typology of arabic dialects: The role of final consonantality. *Journal of Arabic and Islamic studies*, 9:82–109.

Charles F Ferguson. 1959. Diglossia. *Word*, 15(2):325–340.

William Fisher. 1996. Tsylb syllabification package. FTP archive.

Abdellah Fourtassi, Benjamin Börschinger, Mark Johnson, and Emmanuel Dupoux. 2013. Why is English so easy to segment? In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*, pages 1–10, Sofia, Bulgaria. Association for Computational Linguistics.

Bent Fuglede and Flemming Topsoe. 2004. Jensen-shannon divergence and hilbert space embedding. In *International symposium onInformation theory, 2004. ISIT 2004. Proceedings.*, page 31. IEEE.

Luke Galea and Adam Ussishkin. 2018. *Onset clusters, syllable structure and syllabification in Maltese*. Language Science Press.

John Goldsmith. 2011. The syllable. *The handbook of phonological theory*, pages 164–196.

Ivo Grosse, Pedro Bernaola-Galván, Pedro Carpena, Ramón Román-Roldán, Jose Oliver, and H Eugene Stanley. 2002. Analysis of symbolic sequences using the jensen-shannon divergence. *Physical Review E*, 65(4):041905.

Jack Halpern. 2009. Word stress and vowel neutralization in modern standard arabic. In *2nd International Conference on Arabic Language Resources and Tools*, pages 1–7. Cairo.

Rym Hamdi, Salem Ghazali, and Melissa Barkat-Defradas. 2005. Syllable structure in spoken arabic: a comparative investigation. In *Eurospeech—9th European Conference on Speech Communication and Technology*.

Bruce Hayes. 2009. *Introductory Phonology*. Number 23 in Blackwell Textbooks in Linguistics. Wiley-Blackwell.

Jeffrey Heath. 2020. Moroccan arabic. *Arabic and contact-induced change*, 1:213.

Hla Hla Htay and Kavi Narayana Murthy. 2008. Myanmar word segmentation using syllable level longest matching. In *Proceedings of the 6th Workshop on Asian Language Resources*.

Peter W Jusczyk and Carolyn Derrah. 1987. Representation of speech sounds by young infants. *Developmental Psychology*, 23(5):648.

Daniel Kahn. 2015. *Syllable-based generalizations in English phonology*. Routledge.

Niamh Kelly. 2021. Syllable weight, vowel length and focus in lebanese arabic. *Glossa: a journal of general linguistics*, 6(1).

Salam Khalifa, Abdelrahim Qaddoumi, Jordan Kodner, and Owen Rambow. 2025. Learning cross-dialectal morphophonology with syllable structure constraints. *VarDial 2025*, page 157.

Paul Kiparsky. 2003. Syllables and moras in arabic. *The syllable in optimality theory*, 147:182.

Jordan Kodner. 2016. Simple Syllabify.

Constantine Lignos. 2011. Modeling infant word segmentation. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 29–38, Portland, Oregon, USA. Association for Computational Linguistics.

Jianfei Ma, Zhaoxin Feng, Emmanuele Chersoni, Huacheng Song, and Ziqi Zhang. 2025. Phonothink: Improving large language models' reasoning on chinese phonological ambiguities. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19018–19033.

Stefano Manfredi and Sara Petrollino. 2013. Juba Arabic. *The survey of pidgin and creole languages*, 3:54–65.

Yannick Marchand, Connie R Adsett, and Robert I Damper. 2007. Evaluating automatic syllabification algorithms for english. In *Proceedings of the 6th ISCA Workshop on Speech Synthesis*.

John J McCarthy. 1979. On stress and syllabification. *Linguistic inquiry*, 10(3):443–465.

Christian M Meyer and Iryna Gurevych. 2012. *Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography*. na.

Catherine Miller. 2006. Juba arabic. In Kees Versteegh, editor, *Encyclopedia of Arabic Language and Linguistics*, volume 2, pages 517–525. Brill, Leiden.

Ann M Peters. 1983. *The units of language acquisition*. Cambridge University Press.

Hamdi A Qafisheh. 1977. *A Short Reference Grammar of Gulf Arabic*. ERIC.

Belal A. Rakhieh. 2009. *The Phonology of Ma'ani Arabic: Stratal or Parallel OT*. Ph.D. thesis, University of Essex.

Mike Rosner and Claudia Borg. 2022. D1.25: Report on the Maltese language. Deliverable D1.25 (Public), dated 28-02-2022.

Karin C Ryding. 2005. *A reference grammar of modern standard Arabic*. Cambridge university press.

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

Jonathan Sakunkoo and Annabella Sakunkoo. 2025. Lost and found: Computational quality assurance of crowdsourced knowledge on morphological defectivity in wiktionary. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 998–1003.

Natalie Schrimpf and Gaja Jarosz. 2014. Comparing models of phonotactics for word segmentation. In *Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM*, pages 19–28, Baltimore, Maryland. Association for Computational Linguistics.

Elisabeth Selkirk. 1984. On the major class features and syllable theory. *Language sound structure*.

Jason Shaw, Adamantios I Gafos, Philip Hoole, and Chakir Zeroual. 2009. Syllabification in Moroccan Arabic: evidence from patterns of temporal stability in articulation. *Phonology*, 26(1):187–215.

Winston Wu and David Yarowsky. 2021. On pronunciations in wiktionary: Extraction and experiments on multilingual syllabification and stress prediction. In *Proceedings of the 14th Workshop on Building and Using Comparable Corpora (BUCC 2021)*, pages 68–74.

Tatu Ylönen. 2022. Wiktextract: Wiktionary as machine-readable structured data. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 1317–1325, Marseille, France.