

Not too coarse, not too fine: Morphological generalization over vocalized templates in Moroccan Arabic

Ali Nirheche

anirheche@umass.edu

University of Massachusetts Amherst

Research Article

Keywords: morphological generalization, analogy, templatic morphology, Moroccan Arabic, broken plurals, sound plurals, Generalized Context Model

Posted Date: October 6th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10962489/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Not too coarse, not too fine: Morphological generalization over vocalized templates in Moroccan Arabic

Abstract

Analogue models of morphological generalization predict the inflected form of a novel item from stored lexical items that resemble it, but differ in how similarity is defined and at what level of phonological abstraction analogy operates. Semitic morphology, with its independent consonantal and vocalic tiers, offers an unusually clean testing ground. This paper asks which level of representation—the consonantal root, the abstract CV template, or the vocalized template—best accounts for plural generalization in Moroccan Arabic, and whether fine-grained segmental similarity contributes anything beyond it. Using a corpus of 3,735 singular–plural pairs, I implement a hierarchy of Template Match models and evaluate them by 10-fold cross-validation. The Vocalized Template match model outperforms both the Abstract Template match model and the Root match model. Moving toward a fine-grained segmental representation does not improve performance: a Generalized Context Model restricted by the vocalized template never exceeds Template Match. Decomposing the similarity metric shows why segmental similarity still helped in previous work on Modern Standard Arabic: under abstract restriction, a metric confined to vowels recovers almost all of the vocalic information the template leaves unspecified, while one confined to consonants leaves accuracy unchanged. Morphological analogy in Moroccan Arabic therefore operates at an intermediate level of abstraction, and weighting all consonant and vowel features equally is inappropriate for systems in which consonantal and vocalic tiers carry different grammatical functions.

Keywords: morphological generalization; analogy; templatic morphology; Moroccan Arabic; broken plurals; sound plurals; Generalized Context Model

1 Introduction

Morphological generalization—the ability to extend inflectional patterns to novel or unfamiliar words—is a recurrent theme in both cognitive science and computational linguistics. When speakers encounter an unknown word, they must determine which inflectional pattern to apply, drawing on their knowledge of existing words in the language. This process has been studied extensively through nonce word experiments, beginning with Berko’s (1958) classic wug test for English, and has since been extended to a wide range of languages, including German (Marcus et al., 1995; McCurdy, 2024), Turkish (Becker et al., 2011), Arabic (Dawdy-Hesterberg, 2015; Nirheche, 2025), Dutch (Ernestus and Baayen, 2003; Keuleers et al., 2007), Maltese (Nieder et al., 2021), and Russian (Linzen et al., 2013; Becker and Gouskova, 2016). A consistent finding across this literature is that speakers’ generalizations are gradient rather than categorical: they do not simply apply one dominant pattern to all novel forms, but instead show sensitivity to the statistical patterns of the lexicon, producing less frequent patterns at rates that reflect their distribution among existing words.

One influential approach to modeling this behavior is analogy, in which a speaker predicts the inflectional pattern of a novel form by reference to stored lexical items that resemble it (Nosofsky, 1986; Skousen, 1989; Daelemans et al., 1994; Nakisa et al., 1998; Albright and Hayes, 2003; Dawdy-Hesterberg and Pierrehumbert, 2014). One widely used implementation is the Generalized Context Model (GCM; Nosofsky, 1986), which weights stored exemplars by their phonological similarity to the novel item, allowing both type frequency and segmental similarity to influence predictions. Analogical models of this kind have been applied to English past tense (Alegre and Gordon, 1999; Albright and Hayes, 2003; Rácz et al., 2020), German plurals (Nakisa et al., 1998; Marcus et al., 1995; McCurdy, 2024), Dutch plurals (Keuleers et al., 2007), and Modern Standard Arabic (MSA) plurals (Dawdy-Hesterberg and Pierrehumbert, 2014; Dawdy-Hesterberg, 2015).

A persistent challenge for analogical models concerns how similarity between a novel form and stored items is defined, i.e., at what level of phonological abstraction analogy operates. This question is distinct from the more familiar one of whether morphological generalization is rule-based or analogical at all, which has been pursued computationally for English, German and other systems (Nakisa et al., 1998; Albright and Hayes, 2003). The present study takes analogy as its starting point and asks the further question of what the analogy is computed over. That question is not specific to Arabic: it arises in any morphological system where multiple levels of representation could serve as the basis for generalization, and for Arabic it has been addressed directly by Dawdy-Hesterberg and Pierrehumbert (2014) and Alrashed (2021). Semitic morphology, with its independent consonantal and vocalic tiers, provides a particularly informative testing ground, since the contribution of each tier to morphological generalization can be isolated: an Arabic triconsonantal singular, like *qalam* ‘pen’, can be represented at three progressively more specified levels, each retaining more of the vocalic-prosodic tier than the last. The consonantal root retains only the number of consonants (CCC); the abstract CV template additionally fixes the arrangement of consonant and vowel slots, without specifying their quality (CVCVC); and the vocalized template specifies the vowel qualities as well (CaCaC). Across all three levels the consonantal tier, which carries the word’s core lexical meaning in Semitic morphology, is held constant, so

any difference in how well a level predicts pluralization can be attributed specifically to the vocalic information it adds.

Since the Arabic plural system offers exactly this kind of test case, this system has received considerable attention in computational modeling studies. Plunkett and Nakisa (1997) and Nakisa et al. (1998) modelled MSA plurals using connectionist and nearest-neighbour approaches over full segmental representations, though their concern was with whether a default rule is needed rather than with the level of abstraction at which similarity is computed. Dawdy-Hesterberg and Pierrehumbert (2014) addressed the level question directly, comparing a range of analogical models and concluding that the abstract CV template is the primary level of representation at which morphological generalization occurs. Alrashed (2021), on the other hand, tested the learnability of MSA plurals using machine learning classifiers, and shows that the vocalized template is a better predictor of pluralization, a level of representation that was not tested in the prior-mentioned studies. All previous computational work on Arabic plurals has been conducted on MSA. This is a significant limitation, because MSA is not a natively spoken variety: it is a standardized written and formal register that Arabic speakers acquire through formal education rather than as a first language (Ferguson, 1959; Haeri, 2003). Models trained on MSA data—typically drawn from dictionaries or written corpora—may therefore not reflect the morphological knowledge of native speakers, whose primary linguistic competence is in their spoken variety. Dawdy-Hesterberg and Pierrehumbert (2014) acknowledge this concern, noting that dictionary-based corpora may overrepresent formal or archaic vocabulary that is not part of speakers’ productive knowledge.

Spoken varieties, by contrast, are acquired natively and represent speakers’ primary linguistic system, making them a more direct testing ground for models of morphological learning and generalization. To my knowledge, no previous study has applied such models to the plural system of any spoken Arabic variety. The closest precedent is work on Maltese (Nieder et al., 2021), a natively spoken Semitic language of Arabic descent whose plural system likewise combines sound and broken plural formation. In a production experiment, Nieder et al. (2021) found that Maltese speakers pluralize nonce singulars analogically, on the basis of similarity to existing words, rather than by applying a default rule. Their study, however, was not designed to identify the level of representation over which that analogy operates. Therefore, I use data from Moroccan Arabic plurals in the present study. The plural system of Moroccan Arabic is a challenging testing ground for theories of morphological generalization. Like MSA and other Arabic varieties, it combines two morphological strategies: concatenative sound plurals formed by suffixation and non-concatenative broken plurals formed by internal modification of the singular stem. Unlike MSA, however, the Moroccan system is heavily fragmented: the singular form provides less information about the plural than it does in MSA, and speakers must navigate a large number of competing plural patterns for a given singular form (Author, to appear).

The present paper investigates the level of phonological abstraction at which analogical generalization operates in Arabic, using the plural system of Moroccan Arabic as a test case. Four levels of representation are evaluated as potential bases for analogy: the consonantal root, the abstract CV template, the vocalized template, and full segmental representation, in which consonantal identity also counts. The main argument is that analogy operates at

an intermediate level—more abstract than full segmental representation, but less abstract than previous proposals in which vocalic melody (i.e., vowel quality) plays no role, namely the abstract CV template. Specifically, I argue that vocalized templates drive morphological generalization in Moroccan Arabic. This finding has implications for analogical models of morphology more broadly, since it shows that the standard GCM assumption of weighting all segmental features equally is inappropriate for systems in which consonantal and vocalic tiers carry different grammatical functions.

Section 2 introduces the Moroccan Arabic plural system and reviews the debate over the level of representation at which morphological generalization operates in Arabic. Section 3 presents the corpus of 3,735 singular–plural pairs used throughout the paper and documents the distributional properties that make the system difficult to generalize. Section 4 implements and evaluates analogical models that vary in the level of abstraction of the singular representation used as the basis for analogy. Section 5 asks the complementary question of whether moving below the template level, toward fine-grained segmental similarity that includes consonantal identity, provides any further improvement. Section 6 discusses the implications of these findings for analogical models of morphology, and concludes.

2 Background

This section provides the background needed for the modeling that follows. Section 2.1 describes the plural system of Moroccan Arabic. Section 2.2 reviews the evidence addressing which level of phonological representation serves as the basis for morphological generalization in Arabic. Section 2.3 shows how the present study adjudicates between these positions.

2.1 The Moroccan Arabic plural system

The plural system of Moroccan Arabic, like that of MSA and other Arabic varieties, contains two types of plurals: sound plurals and broken plurals. The former are formed through suffixation, while the latter are formed non-concatenatively through internal modification of the singular stem. This dual system of pluralization is a common feature of Semitic morphology and represents one of the most complex aspects of Arabic grammar.

Sound plurals are formed by suffixation. There are three sound plural suffixes (Harrell, 1962), shown in (1). The suffix *-in* is used primarily with plain adjectives, participles, nouns designating profession or origin, diminutive adjectives, and ordinal numbers; *-a* is less common and appears with nouns designating professional or habitual activities; and *-(a)t*, the most widely distributed of the three, is used with all diminutive nouns, all nouns of unity derived from collectives, and feminine participles used as nouns.

- (1) Sound plurals:
 - a. Suffix *-in*:

zwin	zwin-in	‘handsome’
talət	talt-in	‘third’

- b. Suffix *-a*:
 bənnaj bənnaj-a ‘mason’
 kəddab kəddab-a ‘liar’
- c. Suffix *-(a)t*:
 bnita bnit-at ‘little girl’
 hsab hsab-at ‘calculation’

Broken plurals are formed through internal stem modification rather than suffixation. Harrell (1962) identified approximately 40 broken plural patterns in the language, though they vary considerably in productivity and frequency of use. Following Al Ghadi (1990), these can be divided into major and minor patterns. The major patterns are given in (2). The pattern *CCaCəC* is often used for singulars with four consonants, or with three consonants and a full vowel ([a], [u], [i]). *CCaC* is used for singulars with triconsonantal roots or with two consonants and a full vowel, and covers around half of singulars with the pattern *CXC* (where X is either a consonant or a vowel). *CCaCi* is largely reserved for feminine nouns marked with the suffix *-a*, covering roughly half of singulars with the pattern *CəCCa*. The pattern *CCuC* is also a major pattern that is frequently extended by a final [a]; the status of this vowel is contested (Heath, 1987; Ratcliffe, 2002; Nirheche and Becker, 2025), and it is optional in some forms.

(2) Major broken plural patterns:

- a. *CCaCəC*:
 fəndəq fnadəq ‘hotel’
 blas^fa blajəs^f ‘place’
- b. *CCaC*:
 bənt bnat ‘girl, daughter’
 bir bjar ‘well’
- c. *CCaCi*:
 ʒərda ʒradi ‘garden’
 rəkba rkabi ‘knee’
- d. *CCuC(a)*:
 dərb drub(a) ‘street’
 ktab ktub(a) ‘book’

The minor broken plural patterns are less frequent. They include *CəCCaC* (e.g., *taʒər* ‘merchant’ → *təʒʒar*), *CiCan* (e.g., *bab* ‘door’ → *biban*), *CəCCan* (e.g., *blad* ‘country’ → *bəldan*), *CuCəC* (e.g., *b^fjəd^f* ‘white’ → *b^fujəd^f*), and *CuCaCa* (e.g., *d^fʕif* ‘poor’ → *d^fuʕafa*).

2.2 The level of representation for morphological generalization in Arabic

A major question in Arabic morphology concerns the level of abstraction at which morphological generalization operates. Four levels of representation have been proposed or are

relevant to this question: the consonantal root (specifying consonant slots only, e.g., *CCC*), the abstract CV template (specifying consonant and vowel slots but not vowel quality, e.g., *CVCVC*), the vocalized template (specifying vowel quality in addition to vowel slots, e.g., *CuCiC*), and the full segmental representation (specifying consonant identity as well). The present study investigates which of these levels best captures morphological generalization in the Moroccan Arabic plural system.

2.2.1 The root

The root level has been argued to serve as the base for broken plural formation in Moroccan Arabic by previous theoretical analyses (Al Ghadi, 1990; Nirheche, 2021). These studies reject the singular-based approach adopted for MSA, arguing that the unique phonological and prosodic properties of Moroccan Arabic complicate the mapping between singular and plural forms, and that the root therefore provides a simpler and more unified derivational base. Under this view, singulars with varying prosodic templates (e.g., *CCiCa* and *CaCəC*) that map onto the same broken plural pattern are handled by a single root-level rule (*CCC*). Despite its usefulness for deriving plural forms, the root level presents a significant challenge for predictability: a given root can in principle map onto multiple competing broken plural templates, creating a one-to-many mapping problem that makes morphological generalization difficult. No computational or experimental study has tested the root as a predictor of plural generalization in either MSA or spoken Arabic varieties. The present study therefore evaluates the contribution of root shape, in the simplified form set out in Section 3.4.

2.2.2 The abstract CV template

Traditional root-and-pattern and templatic approaches stress the importance of the CV template as an abstract representation specifying consonant and vowel slots but not vowel quality (McCarthy, 1979, 1981; Hammond, 1988; McCarthy and Prince, 1990). This abstract view has been influential in computational work on MSA plurals: Dawdy-Hesterberg and Pierrehumbert (2014) take the abstract CV template as the basis for restricting the set of items over which analogy is computed, and find it to be the most successful level for predicting plural class. This suggests that the abstract CV template may be sufficient for capturing learnability and generalization of Arabic morphology.

Complementary evidence comes from lexical processing studies. Boudelaa and Marslen-Wilson (2004) conducted a series of masked and cross-modal priming experiments on MSA verbs to test the cognitive reality of the abstract CV template (or CV skeleton) and of the vocalic melody. In the Vocalic Melody condition, primes and targets shared the same vowel sequence (e.g., *a-a-a*) but had different abstract CV skeletons (for example, *?affaqa* ‘pity’ paired with *taraka* ‘leave’). In the CV-Skeleton condition, primes and targets shared the same abstract consonant–vowel arrangement (e.g., *CVVCVCV*) but had different vowel sequences (for example, *fuuji?a* ‘be surprised’ paired with *faaraka* ‘participate’). Primes sharing the abstract CV skeleton facilitated target recognition significantly across all experiments, while primes sharing only the vocalic melody consistently failed to produce any significant priming effect. Boudelaa and Marslen-Wilson concluded that the abstract CV template acts as a

functional morphemic unit in lexical access, whereas the specific vocalic melody does not play an independent role.

2.2.3 The vocalized template

The role of the vocalized template—templates that specify vowel quality (e.g., *CaCaC* rather than *CVCVC*)—remains controversial. While some studies have established that the singular CV template is the main predictor for pluralization in MSA (McCarthy and Prince, 1990; Dawdy-Hesterberg and Pierrehumbert, 2014; Dawdy-Hesterberg, 2015), the importance of vowel quality has received little attention. The limited work that does address this question has produced conflicting results, with disagreement over whether vowel quality plays a significant role in lexical processing (Boudelaa and Marslen-Wilson, 2004) and in learnability and generalization (Alrashed, 2021); evidence on the status of the stem as a unit of access (Idrissi et al., 2025) is also relevant, though as discussed below it does not bear on the contrast directly. For spoken Arabic varieties such as Moroccan Arabic, no previous work has tested the role of either abstract CV templates or vocalized templates in lexical processing or learnability.

Evidence from computational modeling of MSA plurals supports the importance of vowel quality in learnability and generalization. Alrashed (2021) tested the learnability of MSA plurals using machine learning classifiers (*K*-Nearest Neighbors and Support Vector Machines) trained on different levels of phonological representation. Models based only on the abstract CV template achieved approximately 52% accuracy in predicting unseen plural forms. Adding vocalic melody information—specifying the quality of vowels rather than just their slots—improved accuracy significantly, by roughly 8%. Alrashed concluded that for plural formation the specific quality of vowels is a critical cue for generalization, and that an abstract CV template alone is insufficient to account for the patterns in the data.

One further study is sometimes taken to bear on this question is Idrissi et al. (2025), who conducted visual masked priming experiments in MSA comparing the priming effects of the stem and the consonantal root. Their design manipulated the relationship between prime and target across four conditions: primes sharing an inflectional stem with the target (*na-ʕtamid* paired with *ja-ʕtamid*), primes sharing a derivational stem (*mu-ʕtamid* paired with *ja-ʕtamid*), primes sharing only the consonantal root (*ʕtimaad* paired with *ja-ʕtamid*), and phonologically similar but morphologically unrelated primes. Priming was graded across these conditions, and Idrissi et al. argue for a hybrid model in which both the root and the stem serve as units of lexical access, with the stem being the more salient of the two. Two caveats are relevant here. First, the difference between the derivational stem condition and the root condition (the comparison that isolates stem sharing from lexical identity) was marginal. Second, their stems differ from their roots in vowel quality, prosodic structure and affixal consonants simultaneously, and the design includes no condition corresponding to the abstract CV template. Their results therefore do support the stem as a unit of lexical access, but do not by themselves distinguish between the abstract and vocalized levels of representation investigated here.

2.2.4 Fine-grained segmental similarity

A fourth possibility is that analogy operates over no template at all, and that similarity is computed directly over the full segmental representation of the singular, so that consonant identity contributes alongside vowel identity. This is the standard assumption in analogical modelling outside Semitic. Exemplar and memory-based models of inflection typically represent each stored item as a string of segments and compute similarity as an edit distance in which substitution costs are weighted by the featural or natural-class similarity of the segments involved (Skousen, 1989; Nakisa et al., 1998; Albright and Hayes, 2003). Nothing about the level of abstraction needs to be stipulated in advance on this view: whatever generalizations exist are expected to emerge from the aggregate behaviour of similar stored items.

Evidence from English suggests that this may be too fine a level to be the whole story. Albright and Hayes (2003) implemented a purely analogical model of the English past tense using exactly this architecture—natural-class-based segmental similarity over an edit-distance alignment—and compared it with a rule-based learner that induces structured environments of varying generality. The analogical model underperformed. It could not capture patterns that require abstract structural characterizations, and it was prone to favouring implausible outputs on the strength of single, highly similar exemplars. Albright and Hayes concluded that speakers generalize on the basis of abstract structural properties rather than gradient segmental resemblance.

For Arabic, two studies bear on this level. Plunkett and Nakisa (1997) trained connectionist classifiers on MSA plurals using representations in which each phoneme is a vector of sixteen binary features, aligned in consonant and vowel slots, so that both vowel and consonant identity is fully available to the model. They report that generalization to novel nouns is predicted more accurately by the similarity structure of the phonological space than by a hybrid rule-plus-associator model. Their concern, however, is with whether a default rule is needed, not with the level of abstraction at which similarity is computed, and they do not compare representations of differing specificity. Dawdy-Hesterberg and Pierrehumbert (2014) is the only study to have examined this level directly, and it reaches a more positive conclusion. Their best-performing model combines the two approaches: the comparison set is first restricted to items sharing the test item’s abstract CV template, and its members are then weighted by their segmental similarity to the test item. This model outperformed template matching alone, which they take as evidence that segmental similarity contributes something the abstract CV template does not encode. The improvement is small, however, and Section 5.4 returns to it: the results reported below suggest that the benefit is attributable not to segmental similarity as such, but specifically to its vocalic component, which recovers information that the abstract CV template leaves unspecified.

2.3 The present study

The conflicting evidence regarding the level of representation for morphological generalization raises questions that the existing literature has not resolved. The discrepancy between Boudelaa and Marslen-Wilson’s (2004) findings and more recent work by Alrashed (2021)

may reflect differences in the morphological processes under investigation—verb versus noun inflection—or differences between lexical access (priming) and morphological generalization. For Moroccan Arabic, no previous work has addressed this question at all. The present study tests systematically which level of abstraction best accounts for plural generalization, using a hierarchy of analogical models that vary in the specificity of the singular representation they operate over.

In analogical approaches to morphological generalization, a novel item takes the inflectional pattern of similar stored items (Skousen, 1989; Albright and Hayes, 2003; Dawdy-Hesterberg and Pierrehumbert, 2014). Models in this tradition differ in how similarity is defined. In Arabic, according to Dawdy-Hesterberg and Pierrehumbert (2014), similarity is treated as categorical. Two questions are therefore addressed. The first concerns the level of abstraction that best predicts plural generalization: restricting analogy to items sharing the same root (i.e., consonant count), abstract CV template, or vocalized template. The second concerns whether moving below the template level—toward fine-grained segmental similarity that includes consonantal identity—provides any further improvement. To address the first question, I implement three Template Match models that vary in the level of abstraction of the singular representation used as the basis for analogy—Root, Abstract Template, and Vocalized Template—and evaluate each via 10-fold cross-validation on a corpus of 3,735 singular–plural pairs. To address the second question, I compare the best-performing Template Match model against a GCM Restricted model, which weights lexical neighbors by their phonological similarity to the test item rather than treating all template-matching items equally.

3 A corpus of Moroccan Arabic plurals

All modeling reported below is trained and evaluated on a single corpus of Moroccan Arabic singular–plural pairs, which is introduced here. Section 3.1 describes how the corpus was compiled. The remaining two subsections document the distributional properties that make this system a challenging case for morphological generalization: the balance of sound and broken plurals (Section 3.2) and the within-word variation affecting a substantial minority of items (Section 3.3).

3.1 Corpus construction

A corpus of Moroccan Arabic plurals of nouns and adjectives was constructed by combining data from two primary sources. The first is the a corpus of Moroccan Arabic plurals (Author, 2025), which contains 1,166 singular–plural noun pairs derived from the Darija Open Dataset (DODa; Outchakoucht and Es-Samaali, 2021). This corpus contains nouns transcribed in IPA and annotated with their corresponding plural forms, English glosses, patterns (or templates), and plural type classification (sound or broken). The second source is Harrell and Sobelman’s (1966) Moroccan Arabic–English dictionary, from which 4,379 items were extracted. The dictionary is well suited to the present purpose. Its entries were elicited from Moroccan Arabic native speakers in Fez, Rabat and Casablanca, and the stated

aim was to cover the core vocabulary of everyday life (Harrell and Sobelman, 1966). For this dictionary-based data, singular and plural forms of nouns and adjectives, together with their glosses, were manually extracted from the dictionary entries. Transliteration into IPA, and the assignment of templates and plural type classifications, were made with Python scripts and then manually checked, with a small number of forms corrected.

After combining both sources, words appearing in both corpora were identified and removed using a Python script to avoid duplication. The resulting combined corpus contains 4,792 unique items, of which 3,820 are nouns (80%) and 972 are adjectives (20%).¹

Since Harrell and Sobelman’s dictionary was compiled in the 1960s, a concern is that some of its entries may reflect vocabulary that is no longer in active use in contemporary Moroccan Arabic. Corpus attestation was therefore used as a filter for identifying which items are still current, and for excluding archaic forms. Each singular and its corresponding plural were looked up in *Atlaset* (Bounhar and Majjodi, 2025), arguably the most comprehensive publicly available collection of Moroccan Arabic text, containing approximately 1.17 million rows of text totaling 1.13 GB of data drawn from a wide variety of sources, including news portals, local forums, cultural blogs, personal narratives, and social media content.

A substantial number of items turned out to be unattested in *Atlaset*, which motivated a manual review of all such items. Of the 1,131 items whose plural was unattested, 1,057 were identified as forms no longer used in contemporary Moroccan Arabic and were removed from the corpus. The remaining 74 unattested items were retained: these are words attested in modern Moroccan Arabic whose plural forms happen not to appear in *Atlaset*, likely because of the narrow contexts in which they are used. The final corpus contains 3,735 singular-plural pairs: 2,982 are nouns (80%) and 753 are adjectives (20%).

3.2 Distribution of plural types

The distribution of plural types reveals several important properties of the Moroccan Arabic plural system. As shown in Table 1, sound plurals constitute the majority of plural types in the corpus (60.6%), while broken plurals account for 37.7% of types. A small proportion of items (1.7%) undergo double pluralization by taking both sound and broken plural inflection at the same time. These are single forms exhibiting both operations simultaneously, i.e., an internal modification of the stem and a plural suffix. Because they involve internal change, they are classified with the broken plurals throughout. Among the sound plural suffixes, *-at* is by far the most frequent by type, accounting for 62.7% of sound plural types. The suffixes *-in* and *-a* are less frequent, representing 26.6% and 10.7% of types, respectively.

¹The corpus, model implementations, and analysis scripts are available at https://osf.io/tvgzy/?view_only=4d8eb9eb20d04f95b2ffc1426862ca39 (anonymized for review).

	Count	%
Overall plural types		
Sound plural	2,264	60.6%
Broken plural	1,409	37.7%
Both sound & broken	62	1.7%
Total	3,735	100%
Sound plural suffixes		
Suffix <i>-at</i>	1,419	62.7%
Suffix <i>-in</i>	602	26.6%
Suffix <i>-a</i>	243	10.7%
Total	2,264	100%

Table 1: Distribution of plural types by type frequency.

Previous work shows varying perspectives on whether sound plurals constitute a minority of the plural forms in MSA, with some proposals suggesting that broken plurals constitute the majority (McCarthy and Prince, 1990; Boudelaa and Gaskell, 2002) and others showing that sound plurals are the majority (Dawdy-Hesterberg and Pierrehumbert, 2014; Sabri et al., 2020). These differences follow from the different sources each researcher consults. Studies using dictionaries and older sources are more likely to find broken plurals to be statistically dominant, while those using more recent online corpora find sound plurals to be dominant. It is worth noting that, even though most type frequency data in the present corpus comes from a dictionary (Harrell and Sobelman, 1966), the sound plural is still the dominant pattern in Moroccan Arabic.

Figure 1 extends the summary in Table 1 by breaking down the broken plural category into its individual patterns, showing the 15 most frequent patterns by type frequency. The sound plural *-at* is the single most frequent pattern in the corpus, with 1,419 types (38%). Among the broken plurals, *CCaCəC* is the most frequent pattern (580 items; 15.5%).

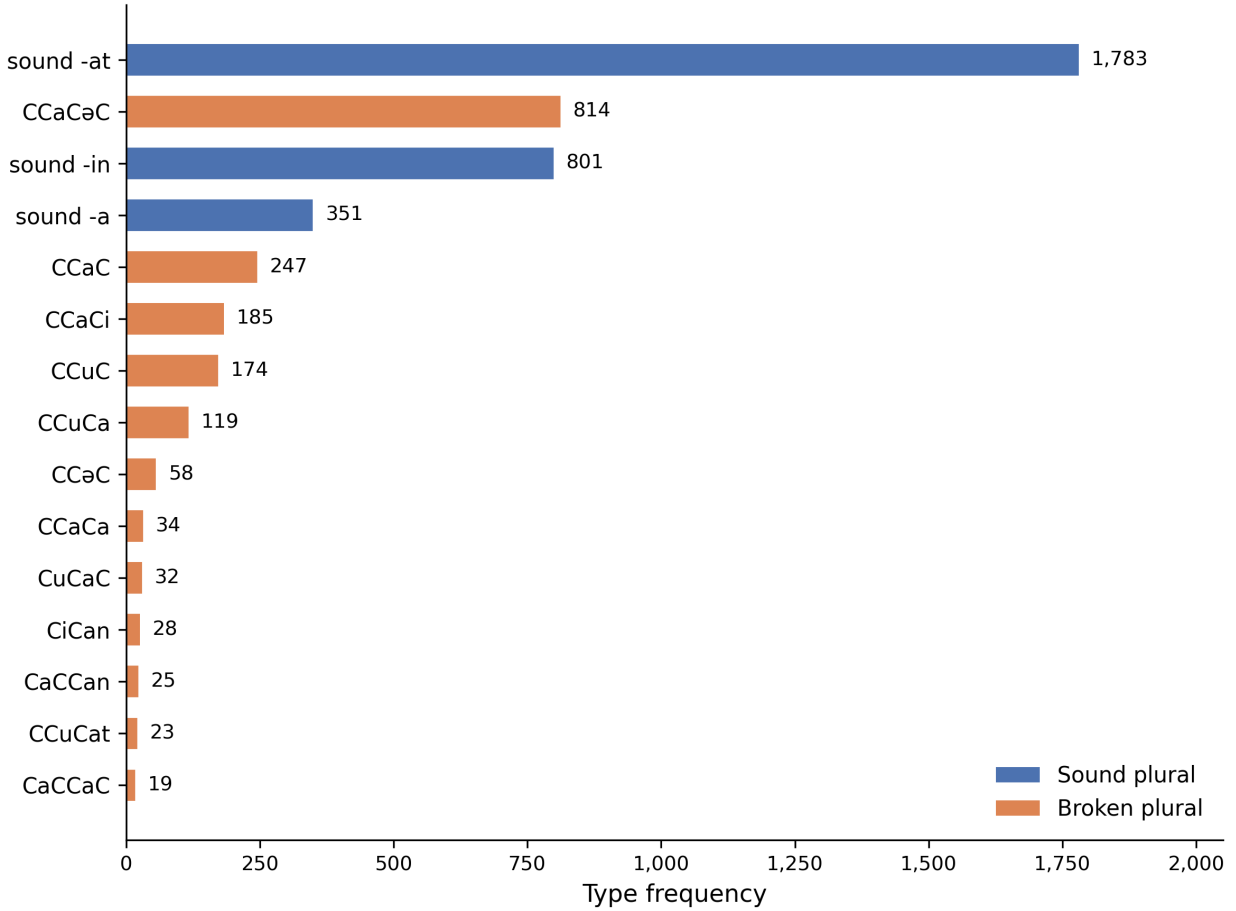


Figure 1: The 15 plural patterns with the highest type frequency in the corpus

3.3 Within-word variation

One major source of complexity in the Moroccan Arabic plural system is within-word variation: cases where a single singular item can take multiple plural forms. For example, the noun *dəʕwa* ‘prayer’ may be pluralized as either *dʕawi* or *dəʕwat*. The corpus contains 3,153 distinct singulars; among these, 482 (15.3%) exhibit this type of variation. Nouns show a higher tendency for variation, with 424 out of 2,494 (17%) taking multiple plural forms, compared to adjectives, where 58 out of 659 (8.8%) do so. As shown in Table 2, the majority of variable singulars (432) have two plural forms, though some take up to four.

Num. of plurals	All singulars	Nouns only	Adjectives only
2	432	384	48
3	47	38	9
4	3	2	1

Table 2: Distribution of the number of attested plurals per singular, over the 482 singulars showing variation.

3.4 Representing the templatic levels in the corpus

Applying the three templatic levels of representation to the corpus requires deciding how much segmental detail each level keeps. Two cases need special consideration: the schwa, which is retained at both template levels, and consonant quality, which is retained at neither.

Templates in the Semitic tradition are not composed exclusively of root consonants and melodic vowels. They often contain fixed segmental material belonging to neither tier: the nominal *m-* of MSA *maktab* ‘office’ and the *ʔa-* of *ʔazraq* ‘blue’ are specified as part of the pattern itself, on a tier distinct from both the consonantal root and the vocalic melody (McCarthy, 1979, 1981). Abstracting away from the vocalic melody therefore does not entail abstracting away from every segment that occupies a vowel position, and abstracting away from the root does not entail abstracting away from every consonant.

Schwa. Moroccan Arabic schwa differs in status from the full vowels [a], [i], and [u]: it is prosodically conditioned and restricted in its distribution in ways that the full vowels are not (Benhallam, 1980; Al Ghadi, 1990; Boudlal, 2001; Bensoukas and Boudlal, 2012; among others). I therefore treat schwa as standing outside the vocalic melody, in a manner comparable to non-root consonants of *maktab* and *ʔazraq* in MSA. Although the latter is morphologically specified and the former phonologically conditioned, it is still the case that neither belongs to the tier that the abstract level is designed to leave unspecified. Schwa is accordingly retained at both the abstract and the vocalized level, which differ only in whether the quality of the full vowels is specified, as shown in (3).

(3) The three levels for two singulars:

- a. *ktab* ‘book’:
 - Root (consonant count) *CCC*
 - Abstract template *CCVC*
 - Vocalized template *CCaC*
- b. *fəndəq* ‘hotel’:
 - Root (consonant count) *CCCC*
 - Abstract template *CəCCəC*
 - Vocalized template *CəCCəC*

The segmental categories of the abstract level are therefore {C, ə, V}, while those of the vocalized level are {C, ə, a, i, u}. This decision is consequential rather than notational.

Table 3 gives the plural distributions of the two abstract templates that a {C, V} skeleton would collapse into a single class. Nouns with the template $CC\textcircled{a}C$ take a broken plural in 128 of 129 cases (99%), with $CCuCa$ the most frequent outcome; nouns with the template $CCVC$ take the sound plural *-at* in 57% of cases and a broken plural in only 43%. Collapsing schwa with the full vowels would merge these 244 nouns into a single class whose most frequent plural is the majority pattern of neither subclass.

Abstract template	Plural pattern	Count	%
$CC\textcircled{a}C$ ($n = 129$)	$CCuCa$	49	38
	$CCaC$	34	26
	$CCuC$	31	24
	$CCuCat$	8	6
	$C\textcircled{a}CCan$	3	2
	other	4	3
$CCVC$ ($n = 115$)	sound <i>-at</i>	65	57
	$CCaC\textcircled{a}C$	21	18
	$C\textcircled{a}CCan$	9	8
	$CCuCa$	4	3
	$CCuC$	3	3
	other	13	11

Table 3: Plural distributions of the two abstract templates that a {C, V} skeleton would collapse.

Consonant quality. The opposite decision is taken for root and singular template consonants, which are represented as undifferentiated slots at all three levels, including where they are plausibly non-root, as in the initial *m-* of *mḥənqa* ‘gallows’ or *mənzila* ‘status’. The primary reason is one of design: holding consonant quality constant across all three levels is what makes the hierarchy a test of the vocalic-prosodic tier specifically, since any difference between levels can then be attributed to the vocalic information added rather than to a confound with consonantal specification. The independent contribution of consonantal identity is not assumed away but tested separately, in Section 5.

The same criterion applied to schwa above also supports this treatment. Specifying the identity of a non-root consonant would subdivide an existing template class only where the resulting cells differ in their plural distribution, and for the relevant cases they do not. In some instances, there is no subdivision: every noun with the abstract template $CC\textcircled{a}CCa$ is *m*-initial, so specifying that consonant yields a single cell, i.e., a model identical to the one that does not. In others, the subdivision is neutral: $m\textcircled{a}CC\textcircled{a}C$ and $C\textcircled{a}CC\textcircled{a}C$ map alike to $CCaC\textcircled{a}C$, so distinguishing them partitions the class without changing any prediction. Retaining schwa and abstracting over consonant quality therefore reflect a single criterion rather than two conflicting ones: a distinction is encoded in the representation when it

partitions a template class into cells with divergent plural distributions, and abstracted over when it does not.

The root. The root is represented in simplified form: each consonant of the singular is taken to correspond to one root consonant, so that the root level reduces to consonant count. Richer definitions are available. Some analyses treat certain consonants as extra-radical and exclude them from the root, such as the prefixal *m-* of *mʃənqa*; others admit vocalic material into the root itself, resulting in shapes such as *CVC*, *CVCC*, and *CCVC* alongside the purely consonantal ones.² Neither is adopted here. Assigning a root of either kind requires item-by-item annotation that cannot be derived automatically from the surface form, and, as will be discussed in Section 4.4, the choice is unlikely to alter the position of this level within the hierarchy.

A further point should be made explicit. What is evaluated here is root shape, not root identity, and the two come apart in a generalization setting. A theory in which items sharing a root share a plural cannot be tested by predicting held-out items, because the root of a held-out item will typically not recur in the training data at all. Root identity is therefore not a candidate basis for generalization to unfamiliar forms, whatever its merits as a basis for derivation. The root level is included here as the coarsest shape-based restriction against which the two template levels can be compared, and the results below should be read accordingly.

4 The role of the vocalized template

This section investigates which level of abstraction in the singular representation best predicts pluralization. Section 4.1 introduces three Template Match models that differ only in that respect—Root, Abstract Template, and Vocalized Template—together with a baseline model. Section 4.2 specifies the mechanisms these models share, and the cross-validation procedure used to evaluate them. Section 4.3 reports the results, and Section 4.4 interprets them and discusses the implications for the debate reviewed in Section 2.2.

4.1 Template Match models

I implemented three Template Match models that differ in the level of abstraction of the singular representation used as the basis for analogy, illustrated in (4) for the noun *munika* ‘doll’ (plural *munikat*).

- (4) a. **Root match model:** Items are considered similar if they share the same consonant count (e.g., *CCC*).
- b. **Abstract Template match model:** Items are considered similar if they share the same abstract CV template, specifying consonant and vowel slots without the quality of the full vowels (e.g., *CVCVCV*).

²See Al Ghadi (1990) and Nirheche (2021) for Moroccan Arabic root shapes containing vowels.

- c. **Vocalized Template match model:** Items are considered similar if they share the same vocalized template, specifying both the slots and quality of the full vowels (e.g., *CuCiCa*).

All models operate on a principle of type frequency: they restrict the training data to words sharing the same template as the test item at the relevant level of abstraction, then select the plural pattern that is most frequent within that restricted set. A Baseline model serves as the lower limit against which all other models are compared; it ignores all template information and simply predicts the globally most frequent plural pattern in the training data, the sound plural *-at*. The segmental content of each level is as described in Section 3.4: schwa is retained at both template levels, and consonant quality at neither, so that any difference between levels reflects the vocalic information added. The independent contribution of consonantal identity is examined separately in Section 5.

Two properties of these representations, motivated in Section 3.4, should be kept in mind. First, schwa is retained at both the abstract and the vocalized level, so the step between them isolates the quality of the full vowels rather than the schwa–full vowel contrast; for a singular whose only vowels are schwas, such as *fəndəq*, the two levels coincide. Second, consonant *quality* is not included as a feature in any model: consonants are treated as abstract placeholders representing only count and positional slot, including where they are plausibly non-root. This reflects the standard assumption in Semitic morphology that the consonantal root encodes lexical-semantic meaning rather than grammatical information, and it is what allows any difference between levels to be attributed to the vocalic information added. The independent contribution of consonantal identity is examined separately in Section 5.

Because the present study includes both nouns and adjectives, and because these two lexical categories may pluralize differently, each of the three Template Match models and the Baseline was implemented in two variants: one that ignores category distinctions (–Category) and one that restricts the comparison set to items in the same grammatical category as the test item (+Category). This results in eight models in total.

4.2 How the models work

The Template Match models implement a frequency-based learning strategy through three core mechanisms: categorical restriction, template match restriction, and frequency-based plural selection.

Categorical restriction. The lexical category restriction mechanism filters the training data to include only items matching the test item’s grammatical category (noun vs. adjective). This filtering reflects the hypothesis that nouns and adjectives constitute distinct morphological classes with different pluralization preferences. As a result, models that implement lexical category restriction maintain separate frequency distributions for nouns and adjectives, preventing cross-category generalization, while models without this mechanism treat all training items as equally relevant regardless of their lexical category.

Template match restriction. The template match restriction mechanism further narrows the comparison set to training items sharing the test item’s singular template at the relevant level of abstraction. The Root match model filters by consonant count only (e.g., *CCC*); the Abstract Template match model filters by the full abstract CV skeleton (e.g., *CVCVCV*); and the Vocalized Template match model filters by the fully vocalized template specifying vowel quality (e.g., *CuCiCa*).

Defaulting to global frequency. When template match restriction results in an empty comparison set, the model implements a fallback strategy to ensure complete coverage. This situation arises during cross-validation: if a singular template is represented by only a single item in the corpus, and that item appears in the test set for a given fold, the training set will contain no items with that template, leaving the model with no exemplars to restrict to. In such cases, category-restricted models fall back to the most frequent plural template within the test item’s category, ignoring template restriction, while non-category-restricted models fall back to the globally most frequent plural template across all training items.

This fallback strategy reflects a hypothesis about which pattern serves as the default (or elsewhere case) in Moroccan Arabic pluralization. Following Dawdy-Hesterberg and Pierrehumbert (2014), I refer to the set of lexical items that share a given singular-to-plural template mapping as a *lexical gang*: every noun with the singular template *CCəC* that pluralizes as *CCuCa*, for instance, belongs to one gang, while those that instead pluralize as *CCaC* belong to a separate gang defined by that mapping. Usage-based models of morphological productivity hold that a pattern remains available for extension to novel forms only if it is supported by a sufficiently large gang: a pattern attested in just a small number of lexical items lacks the statistical weight needed to license generalization, and speakers are predicted to fall back on a more general pattern instead (Bybee, 2001; Pierrehumbert, 2001). The fallback mechanism implemented here applies exactly this logic to the limiting case where the singular has no matching gang in the training data at all.

Which pattern serves as that general or default pattern is itself unsettled in the Arabic morphology literature, with no consensus on which plural type constitutes the elsewhere case in MSA (McCarthy and Prince, 1990; Boudelaa and Gaskell, 2002; Dawdy-Hesterberg and Pierrehumbert, 2014; Mashaqba et al., 2020). Early prosodic accounts treat the sound plural as a minority default that applies only when broken pluralization fails (McCarthy and Prince, 1990), but corpus and experimental evidence points to a stronger role: sounds plurals accounts for approximately 75% of nouns in contemporary MSA corpora and is the preferred choice for borrowings (Boudelaa and Gaskell, 2002), and it is regularly overgeneralized by both children and adults acquiring the language (Mashaqba et al., 2020), which is the expected behavior of a default pattern in usage-based morphology. The fallback strategy adopted here extends this proposal to Moroccan Arabic, and further distinguishes the fallback by lexical category, following the category-typical suffix distribution already noted in Section 2.1: *-in* attaches primarily to adjectival forms, while *-(a)t* is the more broadly distributed suffix for nouns (Harrell, 1962). Table 4 summarizes the resulting fallback pattern for each model type.

Model type	Test item category	Fallback plural
Category-restricted	Noun	sound <i>-at</i>
Category-restricted	Adjective	sound <i>-in</i>
Non-category-restricted	—	sound <i>-at</i> (global default)

Table 4: Fallback plural pattern assigned when no training items share the test item’s template, by model type.

Frequency-based selection. Once the comparison set has been restricted through either template matching or fallback, the model selects the plural template with the highest type frequency within that set. Ties are resolved in favor of the candidate with the higher type frequency in the training partition as a whole. Type frequency counts distinct lexical items rather than weighted token frequencies, consistent with evidence that morphological productivity correlates more strongly with type than with token frequency (Baayen, 1992; Pierrehumbert, 2001).³

The candidates over which plural selection is made are not always fully templatic. If a restricted set contains the sound plural templates *CəCCat* and *CəCCaCat* for instance, the two would be merged into a single candidate *-at*, and the same is true for *-in* and *-a*. In contrast, broken plural and doubly pluralized templates remain distinct plural candidates. The reason for defining the candidate set at this level is given below, once the accuracy metric has been introduced.

One convention about the input should also be mentioned here. Singulars with more than one attested plural enter the models as one pair per attested plural, so that each attested plural contributes independently to the type frequency of its pattern; a singular with two attested plurals contributes evidence for both patterns.

Figure 2 shows how the three models differ when applying the template restriction mechanism to the same test item, the noun *zar* ‘neighbor’ (correct plural: *ziran*, template *CiCan*). All three models first restrict the comparison set to nouns, then apply their respective template restrictions, and finally select the most frequent plural pattern within the resulting set. The three models impose progressively narrower template restrictions on the training set, resulting in different comparison sets and therefore different predictions.

³But see Jarosz et al. (2025) for artificial language learning evidence that token frequency also contributes to morphological generalization.

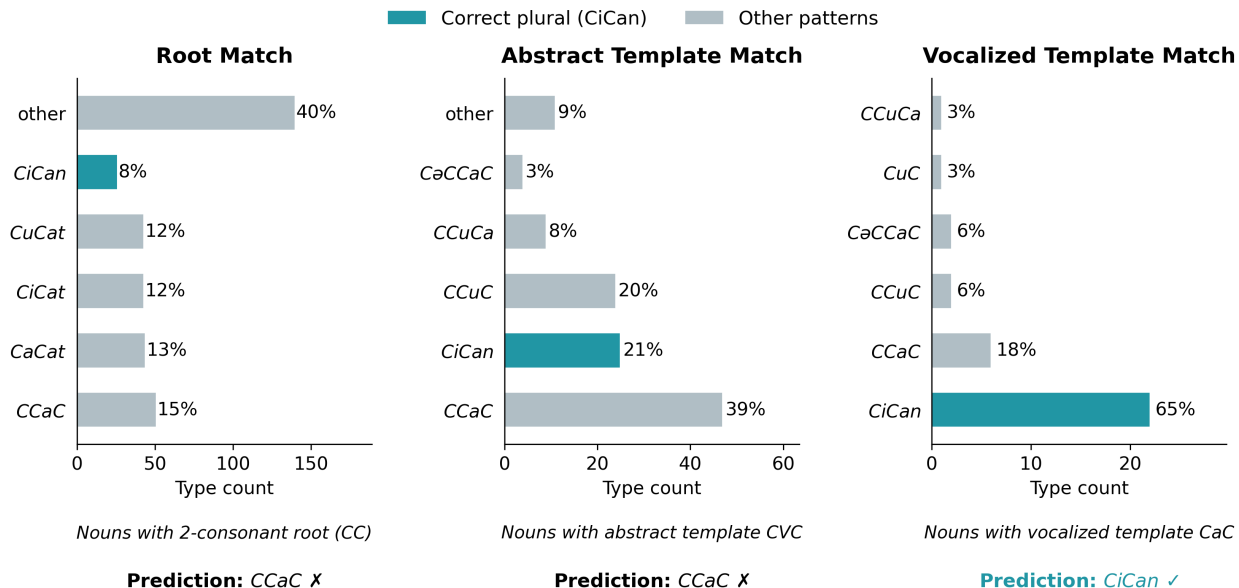


Figure 2: Distribution of plural patterns within the comparison set for each +Category Template Match model, applied to the test noun *zar* ‘neighbor’ (correct plural: *ziran*, pattern *CiCan*). Teal bars indicate the correct plural pattern. Each panel shows the restriction criterion applied and the model’s prediction

The Root match model imposes a very coarse restriction (all 2-consonant nouns) which produces a large comparison set of 347 nouns dominated by the broken plural *CCaC* (15%), with the correct plural template *CiCan* ranking fifth at 8%. The Abstract Template match model narrows this to nouns sharing the abstract template *CVC*, which again promotes *CCaC* to the top (39%); however, since *CVC* collapses forms with different vowel qualities (e.g., *CaC*, *CiC*, *CuC*), the within-set distribution is still spread across multiple competing patterns, and *CiCan* ranks second at 21%. In contrast, by restricting to nouns sharing the vocalized template *CaC*, the Vocalized Template match model isolates the subgroup in which *CiCan* is the majority pattern at 65%, and correctly predicts the plural of *zar*.

Model evaluation. To ensure robust and replicable evaluation of model performance, I used 10-fold cross-validation with standardized fold assignments. The dataset was partitioned into ten non-overlapping folds, with each fold serving once as the test set while the remaining nine folds served as the training set. Crucially, the fold assignments were fixed across all model variants, ensuring that differences in accuracy reflect differences in model architecture rather than variability in data splits.

For each fold, the model was trained on the training partition and evaluated on the held-out test partition. Performance was measured as the proportion of correct predictions, where correctness was defined differently for sound and broken plurals. For broken plurals, a prediction was considered correct only if it exactly matched the target plural template (e.g., *CiCan*), since the template fully determines the non-concatenative stem modification that takes place. For sound plurals, similar to the methodology used by Dawdy-Hesterberg

and Pierrehumbert (2014), a prediction was considered correct if both the predicted and actual forms ended with the same suffix (*-at*, *-in*, or *-a*), regardless of differences in the stem. This is justified by the concatenative nature of sound pluralization: sound plurals are formed by suffixation, and the plural morphology is entirely encoded in the suffix. This is why the candidate set introduced above merges sound plurals by suffix: selection and evaluation must operate at the same degree of specificity, so that a frequent suffix could not lose to a less frequent broken template that happens not to be subdivided. Two consequences follow. First, the doubly pluralized forms are treated as broken plurals: they involve internal modification, and are therefore scored by exact template match rather than by suffix identity. Second, because the models are deterministic and select a single pattern, they cannot predict both members of a singular with two attested plurals: each pair is scored separately against its own attested plural.

Accuracy was computed separately for each fold, and the mean accuracy across all ten folds is reported as the model’s overall performance. Standard deviations and 95% confidence intervals were calculated to assess the stability of predictions across folds.

Relation to previous work. The Template Match models implemented here share fundamental similarities with the Simple Template Match model presented by Dawdy-Hesterberg and Pierrehumbert (2014) in their computational study of MSA plurals. Both approaches restrict the comparison set to words sharing the test item’s singular template and select the most frequent plural pattern within that restricted set. The present study extends this framework in three ways. First, Dawdy-Hesterberg and Pierrehumbert (2014) restricted their model to the abstract CV template (e.g., *CVCVCV*), whereas the current hierarchy systematically tests multiple levels of abstraction from root to vocalized template. This makes it possible to determine empirically which level of representation is most predictive, rather than assuming abstract CV template level a priori. Second, the models presented here incorporate a fallback strategy that handles novel template patterns by reverting to category-level or global frequency distributions, ensuring complete coverage of the test set. The Dawdy-Hesterberg and Pierrehumbert (2014) model lacked such a mechanism, potentially limiting its ability to generalize to rare or unseen template patterns. Third, the models used here systematically manipulate the presence or absence of categorical restriction, isolating the contribution of grammatical category information, a factor not examined in the previous work.

4.3 Results

Figure 3 presents the mean 10-fold cross-validation accuracy of all eight models. Accuracy increases consistently as models gain access to more information about the singular form, and lexical category filtering improves performance at every level of template specification. Table 5 shows the accuracy of all eight models, along with the proportion of test items for which the template restriction failed and the fallback occurred.

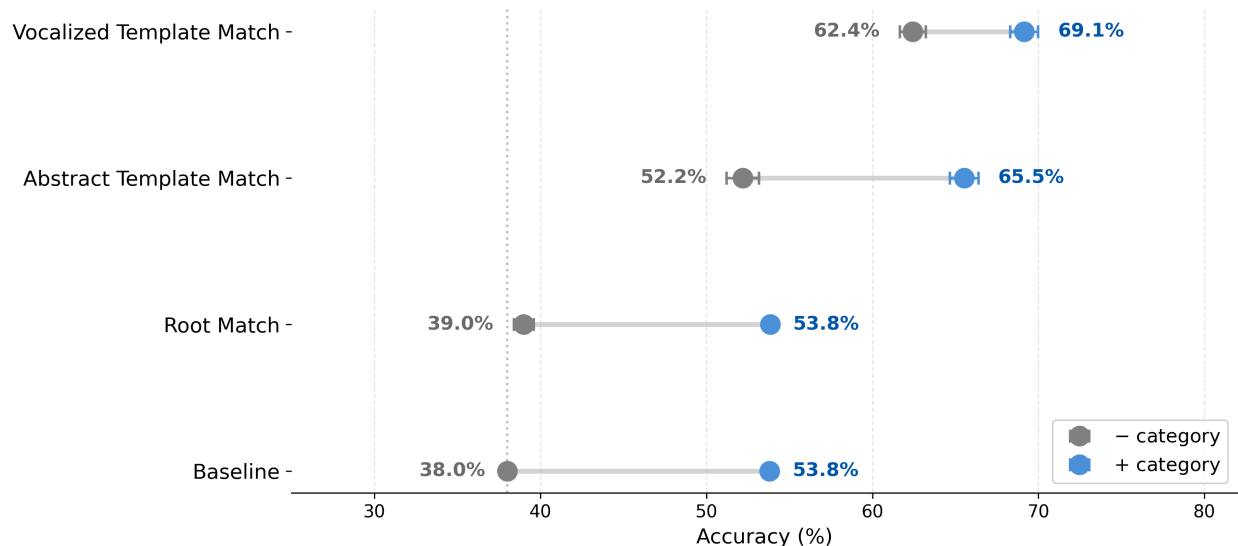


Figure 3: Accuracy across the hierarchy of models. Gray circles represent models without lexical category filtering; blue diamonds represent models with lexical category filtering. Error bars show 95% confidence intervals from 10-fold cross-validation

Model	Accuracy	SD	95% CI	Fallback
<i>– Category</i>				
Baseline	37.99%	0.15%	[37.90, 38.08]	—
Root	38.98%	0.97%	[38.38, 39.59]	0%
Abstract Template	52.18%	1.58%	[51.21, 53.16]	0.70%
Vocalized Template	62.43%	1.26%	[61.65, 63.21]	2.17%
<i>+ Category</i>				
Baseline	53.79%	0.32%	[53.59, 53.99]	—
Root	53.84%	0.28%	[53.67, 54.02]	0%
Abstract Template	65.52%	1.39%	[64.66, 66.38]	0.70%
Vocalized Template	69.13%	1.36%	[68.29, 69.98]	2.17%

Table 5: Mean 10-fold cross-validation accuracy of all eight models, with standard deviations, 95% confidence intervals, and the proportion of test items predicted through the fallback mechanism.

The Baseline model, which ignores all singular information and simply predicts the most frequent plural pattern in the training data, achieves 37.99% accuracy without category information and 53.79% with it. The best-performing model—Vocalized Template match with category filtering—achieves 69.13%, an improvement of 31.14% over the category-blind baseline and 15.34% over the category-aware baseline. Restricting accuracy to the items

decided by template match rather than fallback leaves this figure almost unchanged (68.94%, $n = 3,654$), so the model’s advantage does not derive from the fallback mechanism.

A mixed-effects logistic regression was conducted to examine whether each model predicted each test item correctly. The fixed effects are the level of representation at which the comparison set was restricted and whether the model had access to lexical category; random intercepts are included for item and for target plural pattern. As shown in Table 6, relative to the Baseline, restriction by the root does not improve accuracy ($\beta = 0.07$, $p = .234$), while restriction by the abstract template does ($\beta = 1.69$, $p < .001$), and restriction by the vocalized template does so more strongly ($\beta = 2.59$, $p < .001$). Access to lexical category contributes independently of the level of representation ($\beta = 1.67$, $p < .001$).

Fixed effect	β	SE	z	p
Intercept	−0.833	0.306	−2.72	.007
Root (vs. Baseline)	0.072	0.061	1.19	.234
Abstract template (vs. Baseline)	1.691	0.063	26.88	< .001
Vocalized template (vs. Baseline)	2.588	0.068	38.16	< .001
Category awareness	1.673	0.047	35.62	< .001

Table 6: The mixed-effects logistic regression model results.

Refitting the same model with the abstract template as the reference level gives the contrast between the abstract and vocalized template; the results of this model show that the vocalized template outperforms the abstract template ($\beta = 0.90$, $p < .001$; mean fold-level difference = 3.61%). Specifying vowel quality in the singular representation therefore provides predictive power beyond the abstract CV template alone, even when category information is held constant. The gap between the root and both abstract and vocalized template levels stays larger: relative to the abstract template, restriction by the root is worse by a wide margin ($\beta = -1.62$, $p < .001$), and the fold-level difference between the root and vocalized models is 15.29%. The category effect is estimated across all four levels simultaneously and is significant ($\beta = 1.67$, $p < .001$). At the vocalized template level specifically, the +Category variant exceeds its −Category counterpart by 6.70%.

4.3.1 Accuracy by plural type

Accuracy was also computed separately for sound and broken plurals (Table 7). Two findings are worth mentioning here. First, neither the Baseline nor the Root match model predicts broken plurals at all: both assign a sound plural suffix to every test item. Second, the improvement attributed to template restriction entirely comes from the broken plurals, rising from zero to 24.95% at the abstract level and 35.01% at the vocalized level. The vocalized template’s advantage over the abstract CV template in the category-aware condition does not come from sound plurals (91.30% vs. 91.87%), but exclusively from the broken plurals (35.01% vs. 24.95%). Vowel quality, in other words, does not make plural prediction uni-

formly easier. It does the specific work that the broken plural requires, which is to identify which non-concatenative pattern a given singular shape selects.

Model	−Category		+Category	
	Sound	Broken	Sound	Broken
Baseline	62.68%	0.00%	88.74%	0.00%
Root	64.22%	0.14%	88.74%	0.14%
Abstract Template	74.65%	17.61%	91.87%	24.95%
Vocalized Template	84.36%	28.69%	91.30%	35.01%

Table 7: Accuracy by plural type ($n = 2,264$ sound, $n = 1,471$ broken).

4.4 Discussion

4.4.1 The null effect of consonant count

The Root match model, which filters the training data by consonant count only, is indistinguishable from the Baseline in the category-aware condition (53.84% vs. 53.79%), with no evidence for a significant difference in the mixed-effects model ($\beta = 0.07$, $p = .234$). In the category-blind condition, the difference is larger (0.99%) but it remains much smaller than the gain the abstract template contributes over the Baseline in the same condition (14.19%). Restricting the comparison set by the number of consonants in the singular therefore does not significantly improve the model’s predictions.

This null effect has a straightforward explanation grounded in the distributional properties of the corpus. As shown in Table 8, the dataset contains only six possible root types, and these are highly skewed in frequency: triconsonantal (CCC) and quadriconsonantal (CCCC) roots account for 85.9% of all words in the corpus (3,210 out of 3,735 items), while other root types are rare. Consequently, most test items fall into the CCC or CCCC categories, and restricting the training data by these dominant root types results in comparison sets that are only slightly smaller than the unrestricted baseline. Within these large sets, the most frequent plural pattern dominates overwhelmingly: the sound plural *-at* in the category-blind condition, or the category-specific defaults (*-in* for adjectives, *-at* for nouns) in the category-aware condition. The Root match model’s predictions are therefore determined by the same high-frequency defaults the Baseline uses, and the two models converge.

Root type	Frequency
C	3
CC	391
CCC	2,031
CCCC	1,179
CCCCC	124
CCCCCC	7
Total	3,735

Table 8: Distribution of root types in the corpus.

An objection one may make is that this conclusion depends on the simplified definition of the root adopted in Section 3.4, and that a richer definition (one excluding extra-radical consonants, or admitting vowels into the root) would yield a different result. These finer definitions would certainly redistribute items across root classes, but they would not move this root level higher in the hierarchy. Even the most finer definition of a root would not result in more than ten root shapes, against the 60 abstract CV templates and 187 vocalized templates attested in the corpus, so restriction at the root level would remain far coarser than the abstract and vocalized template levels, and the comparison sets it would produce would remain correspondingly heterogeneous.

4.4.2 The relationship between category and template specification

A consistent pattern across the hierarchy is that the benefit of lexical category filtering declines as template specificity increases. Table 9 shows the accuracy gain from adding category information at each level.

Model	−Category	+Category	Boost (pts)
Baseline	37.99%	53.79%	+15.80
Root	38.98%	53.84%	+14.86
Abstract Template	52.18%	65.52%	+13.34
Vocalized Template	62.43%	69.13%	+6.70

Table 9: Accuracy gain from adding lexical category information at each level of template specification.

The category boost declines from roughly +15% at the Baseline and Root levels to +6.70 at the Vocalized Template level. This pattern reflects the fact that lexical category and vocalized template information are partially redundant: vocalized templates themselves partition nouns and adjectives into largely distinct morphological classes. Table 10 shows that of the vocalized templates in the corpus, 56.1% are used exclusively by nouns, 17.1%

exclusively by adjectives, and 26.7% are shared between categories. Knowing the vocalized template therefore already provides substantial category-like information directly.

Template type	Count	%
Noun-only	105	56.1%
Adjective-only	32	17.1%
Shared	50	26.7%
Total	187	100%

Table 10: Distribution of vocalized templates across lexical categories.

By contrast, at the Abstract Template level, only 28.3% of templates are noun-specific and 11.7% are adjective-specific, with 60% shared between categories (Table 11). This substantially higher degree of overlap explains why explicit category information remains highly valuable when only vowel slots are available: abstract CV templates do not effectively separate nouns from adjectives, so category filtering provides a critical additional constraint.

Template type	Count	%
Noun-only	17	28.3%
Adjective-only	7	11.7%
Shared	36	60.0%
Total	60	100%

Table 11: Distribution of abstract templates across lexical categories.

Crucially, the category boost does not disappear even at the Vocalized Template level. The residual advantage of 6.70% stems directly from the 26.7% of vocalized templates that are shared between nouns and adjectives (Table 10). For words with these shared templates, the model cannot determine plural class from template identity alone, and explicit category information provides the necessary disambiguation. Grammatical category therefore carries information about plural selection that is not recoverable from the vocalized template, rather than only correlating with vocalic patterns.

4.4.3 Vocalized vs. abstract templates

These results directly address the theoretical debate outlined in Section 2.2 regarding the role of abstract CV templates versus vocalized templates in morphological generalization. The relevant comparison is between the Abstract Template match model and the Vocalized Template match model. As shown in Table 12, adding vowel quality to the abstract CV skeleton improves accuracy substantially in both the category-blind and category-aware conditions.

Model	Category	Accuracy	Difference (pts)
Abstract Template	No	52.18%	+10.25
Vocalized Template	No	62.43%	
Abstract Template	Yes	65.52%	+3.61
Vocalized Template	Yes	69.13%	

Table 12: Accuracy of the Abstract Template match vs. the Vocalized Template match model, with and without category filtering.

Without category information, specifying vowel quality improves accuracy by 10.25%; with category information, the improvement is 3.61%, which is a significant improvement ($\beta = 0.90$, $p < .001$). The smaller gain in the +Category condition reflects the partial redundancy between vocalic and categorical information discussed above. As Table 7 shows, the gain is not spread evenly over the corpus. In the category-aware condition the vocalized template does not provide any improvement over the abstract template on sound plurals and provides a higher accuracy by around ten percent on broken plurals. A mixed-effects logistic regression restricted to these two models confirms the asymmetry: the advantage of the vocalized template is roughly twice as large for broken as for sound plurals (interaction $\beta = 0.71$, $p < .001$).⁴ What vowel quality contributes is therefore specifically the ability to identify which broken plural pattern a singular selects.

These findings diverge from the conclusion of Boudelaa and Marslen-Wilson (2004), who argued on the basis of priming experiments that the abstract CV template functions as the primary morphological unit while vowel quality plays no independent role. The present results suggest that this conclusion does not extend to morphological generalization: the vocalized template outperforms the abstract CV template in predicting plural formation. This is consistent with Alrashed (2021) and Idrissi et al. (2025), who showed computationally and experimentally that vocalized templates carry substantial morphological weight.

The discrepancy between Boudelaa and Marslen-Wilson’s findings and the present results likely reflects a difference between the two tasks: lexical access (priming) versus morphological generalization (predicting the plural of a stored or novel form). The abstract CV template may be sufficient as a unit of lexical access, where the task is to recognize a known word, while vowel quality becomes critical when the task requires generalizing a morphological pattern to a new form, which is precisely the situation modeled here, and the situation speakers face when pluralizing an unfamiliar word.

⁴The interaction is estimated on the Abstract and Vocalized models alone. The Baseline and Root models predict no more than two broken plurals out of 1,471, so a model including them has no variation to estimate from in those cells and its coefficients cannot be interpreted.

5 The role of consonantal similarity

The results above establish that analogy in Moroccan Arabic plurals operates over vocalized templates: models that restrict their comparison set to items sharing the test item’s vocalized template reach 69.13% accuracy, substantially outperforming those operating at coarser levels of abstraction. This raises a complementary question: does moving in the opposite direction—toward a *less* abstract representation that includes fine-grained segmental identity—provide any further improvement? Prior analogical work on other languages relies on segmental similarity—shared phonological features of both consonants and vowels—as the primary basis for generalization (Alegre and Gordon, 1999; Daelemans et al., 1994; Ernestus and Baayen, 2003), and it is possible that such similarity provides information beyond what the vocalized template already encodes.

Dawdy-Hesterberg and Pierrehumbert (2014) addressed a related question for MSA, testing whether segmental similarity adds predictive power alongside the abstract CV template. Using variants of the Generalized Context Model (GCM; Nakisa et al., 2000; Albright and Hayes, 2003), which weights lexical neighbors by featural phonological similarity, they compared five models varying in their use of segmental similarity, type frequency (or gang size), and CV-template restriction, as shown in Table 13.

Model	Seg. similarity	Type frequency	CV-template
Single Best Match Unrestricted	✓	×	×
Single Best Match Restricted	✓	×	✓
GCM Unrestricted	✓	✓	×
GCM Restricted	✓	✓	✓
Simple Template Match	×	✓	✓

Table 13: Five analogical models tested by Dawdy-Hesterberg and Pierrehumbert (2014) for MSA.

The best-performing model was GCM Restricted (65.97%), followed closely by Simple Template Match (63.94%), with the difference significant ($t(9) = 5.06$, $p < .01$). Dawdy-Hesterberg and Pierrehumbert concluded that the abstract CV template is the major predictor of pluralization in MSA, while segmental similarity adds a small but significant boost (2.03% over their Simple Template Match model).

In what follows, I test whether segmental similarity provides any similar benefit in Moroccan Arabic, and specifically whether that benefit—if it exists—comes from consonantal or vocalic similarity. This distinction matters because the vocalized template already encodes vocalic information: a model restricted by the vocalized template has already abstracted away from consonantal identity while retaining vowel quality and prosodic structure. Any remaining contribution of segmental similarity to plural prediction must therefore come specifically from consonantal identity. In Semitic root-and-pattern morphology, however, consonants encode lexical-semantic meaning rather than grammatical information, and

plural class membership is determined by the vocalic-prosodic tier. The prediction is therefore that consonantal similarity will provide no reliable information about plural selection beyond what the vocalized template already captures, and that a GCM model using full segmental similarity—consonants and vowels—will not outperform the Vocalized Template match model. I will also argue that the small boost from segmental similarity reported by Dawdy-Hesterberg and Pierrehumbert (2014) is not evidence that segmental similarity is independently useful for generalization, but rather a consequence of their use of the abstract CV template as the restriction criterion.

5.1 The GCM Restricted model

To test whether consonantal similarity provides any predictive power beyond the vocalized template, I implemented the GCM Restricted model, following Dawdy-Hesterberg and Pierrehumbert (2014) with one modification: category restriction is added, so that the comparison set is limited to training items in the same grammatical category as the test item. This addition is necessary because Dawdy-Hesterberg and Pierrehumbert’s MSA corpus contained only nouns, while the present Moroccan Arabic corpus includes both nouns and adjectives. The model proceeds through four steps. First, category restriction filters the training set to items matching the test item’s lexical category (noun vs. adjective). Second, template restriction further narrows the comparison set to items sharing the test item’s vocalized template. For example, for the test noun *zar* ‘neighbor’ (template *CaC*), the restricted set contains only nouns with the singular template *CaC*, which map to the plural patterns *CiCan*, *CCaC*, *CCuC*, *CaCCaC*, *CaCat*, *CCuCa*, and *CCuC*.

Third, segmental similarity is computed between the test form and each stored exemplar in the restricted set. Similarity is based on string-edit (Levenshtein) distance d_{ij} (Kruskal, 1983), where larger distances indicate lower similarity. Dissimilarity is exponentially transformed into a similarity weight g_{ij} :

$$(5) \quad g_{ij} = e^{\left(-\frac{d_{ij}}{s}\right)^p}$$

Here, s is a sensitivity parameter controlling the decay rate and p is an exponent governing the shape of the decay function. Following Dawdy-Hesterberg and Pierrehumbert (2014), p is set to 1 and s is set to 0.3, the best-fit value they report for their MSA data. It is worth noting that 0.3 is a fitted quantity rather than a constant of the model, and that it was fitted to a different Arabic variety; Section 5.5 therefore reports the behavior of the model across the range of s rather than at this value alone. The parameter s has a straightforward interpretation that will matter below. As s decreases, the exponent grows large in magnitude even for small distances, so weight concentrates on the few exemplars nearest to the test form and the model approaches nearest-neighbor behavior. As s increases, g_{ij} approaches 1 for every exemplar, all members of a gang contribute equally, and the gang score becomes proportional to gang size. In this latter extreme, the GCM Restricted model is the same as the Vocalized Template match model.

The string-edit distance incorporates segmental similarity based on shared natural classes (Frisch et al., 2004), such that changes between more similar segments (e.g., /s/ → /z/) are

assigned a lower cost than changes between dissimilar ones (e.g., /s/ $\rightarrow$ /l/). All segments (consonants and vowels) receive equal weighting, following Dawdy-Hesterberg and Pierrehumbert (2014). At this third step, for the candidate gang $CaC \rightarrow CiCan$, for instance, the model computes the similarity of zar to stored exemplars within that gang such as yar ‘cave’, kar ‘bus’, and far ‘mouse’.

Fourth, the overall similarity S_{iC_J} from test form i to gang C_J aggregates weights across all members of the gang, normalized by the total similarity to all gangs in the restricted set (Nakisa et al., 2000; Albright and Hayes, 2003; Dawdy-Hesterberg and Pierrehumbert, 2014):

$$(6) \quad S_{iC_J} = \frac{\sum_{j \in C_J} g_{ij}}{\sum_{k \in \mathcal{C}} g_{ik}}$$

Where $\mathcal{C}$ is the set of all gangs in the restricted comparison set. Each gang score therefore reflects both the summed segmental similarity between the test form and the stored exemplars exhibiting that pattern, and the type frequency of the pattern (since larger gangs accumulate more similarity mass). The model deterministically selects the gang C_J with the highest S_{iC_J} as its prediction.

Crucially, because all items in the restricted set share the same vocalized template as the test item, their Levenshtein distances to the test form vary only as a function of consonant identity. The vowel slots and vowel qualities are identical across all members of the restricted set by construction. The gang score therefore reduces to a function of type frequency modulated by consonantal similarity alone. This makes the GCM Restricted model a direct test of the hypothesis that consonantal similarity provides information beyond what the vocalized template already encodes: if consonantal identity were informative for plural selection, the GCM Restricted model should outperform the Vocalized Template match model.

Model performance was evaluated using the same 10-fold cross-validation procedure and accuracy metric as the Template Match models: exact plural template match for broken plurals and suffix match for sound plurals.

5.2 Results

Figure 4 compares the 10-fold cross-validation accuracies of the GCM Restricted model and the Vocalized Template match model. The Vocalized Template match model achieves the highest mean accuracy (69.13%), while GCM Restricted achieves 66.25%.

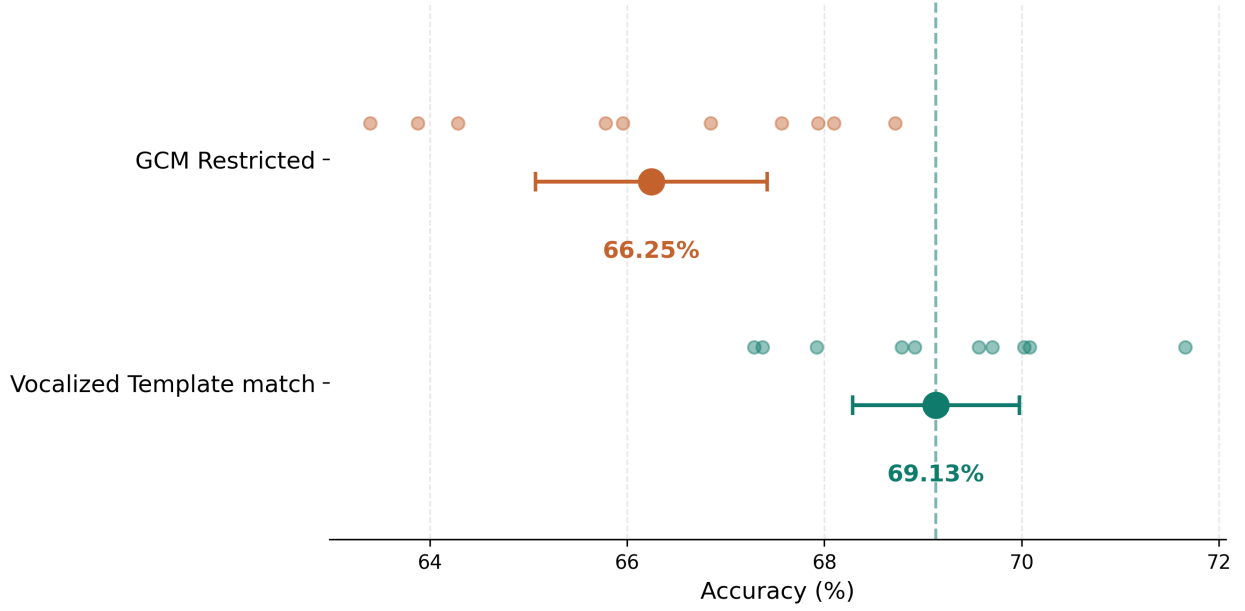


Figure 4: 10-fold cross-validation accuracies for the Vocalized Template match model and GCM Restricted. Colored points show per-fold accuracies; filled circles show means; horizontal lines show 95% CIs

The Vocalized Template match model outperforms GCM Restricted by 2.88%; in a mixed-effects logistic regression of item-level accuracy on the two models, with random intercepts for item and plural pattern, the GCM performs significantly worse ($\beta = -0.59$, $p < .001$). Therefore, at this parameter value, consonantal similarity provides no benefit under vocalized restriction, but it seems to be harmful. An examination of the disagreement between both models shows that the 199 items on which the Vocalized Template match model succeeds and GCM Restricted fails, the single largest group (85 items) consists of feminine nouns (ending in -a) whose plural is the sound plural -at, but the GCM Restricted assigns them the broken pattern *CCaCəC* because minority-pattern training items in the same comparison set happen to share consonantal features with the test item. Table 14 gives examples. The second largest group (29 items) shows the opposite pattern, with GCM Restricted selecting -at where the attested plural is *CCaCəC*.

Singular	Sg. templ.	Templ. Match (correct)	GCM Restr. (incorrect)
<i>bhima</i> ‘beast’	<i>CCiCa</i>	sound -at	<i>CCaCəC</i>
<i>blas^ʃa</i> ‘place’	<i>CCaCa</i>	sound -at	<i>CCaCəC</i>
<i>bluza</i> ‘blouse’	<i>CCuCa</i>	sound -at	<i>CCaCəC</i>
<i>dqiqa</i> ‘minute’	<i>CCiCa</i>	sound -at	<i>CCaCəC</i>
<i>bəs^ʃs^ʃila</i> ‘onion’	<i>CəCCiCa</i>	sound -at	<i>CCaCəC</i>
<i>bət^ʃbut^ʃa</i> ‘duckling’	<i>CəCCuCa</i>	sound -at	<i>CCaCəC</i>
<i>bəlbuzə</i> ‘blister’	<i>CəCCuCa</i>	<i>CCaCəC</i>	sound -at
<i>bərt^ʃma</i> ‘flat’	<i>CəCCa</i>	<i>CCaCəC</i>	sound -at

Table 14: Examples of prediction disagreements between GCM Restricted and the Vocalized Template match model. The upper block shows the largest error group (85 items), and the lower block shows the second largest (29 items).

These results are consistent with the prediction: adding fine-grained consonantal similarity to the vocalized template restriction does not improve plural prediction, and reduces accuracy.

5.3 Neutralizing consonantal similarity

To confirm that consonantal identity is specifically the source of noise in the GCM Restricted model, I implemented a modified version of the model in which all consonants were assigned identical feature specifications, effectively neutralizing consonantal distinctions in the similarity computation. Under this manipulation, the only phonological information available to the model is vowel quality and syllable structure, which is precisely the information encoded in the vocalized template. As shown in Table 15, consonant neutralization improves GCM Restricted accuracy by 2.88% (66.25% $\rightarrow$ 69.13%), matching the Vocalized Template match model exactly.⁵

⁵It should be noted that, because the comparison set has already been restricted to items sharing the test item’s vocalized template, every member of that set has the same vowel qualities and the same arrangement of consonant and vowel slots. Neutralizing consonantal distinctions therefore makes all members of the set indistinguishable from one another, so that $d_{ij} = 0$ and $g_{ij} = 1$ for every exemplar, and the gang score reduces to gang size. The convergence reported below is therefore expected analytically rather than discovered empirically. Its value is that it establishes that the entire gap is attributable to the consonantal component of the similarity metric, with no residual contribution from any other source.

Model	Full features	Consonants neutralized
GCM Restricted ($s = 0.3$)	66.25%	69.13%
Vocalized Template match		69.13%

Table 15: Mean 10-fold cross-validation accuracy of GCM Restricted with and without consonantal feature distinctions, compared to the Vocalized Template match model.

This locates the source of the performance drop in the GCM Restricted model. Because the model has already filtered the comparison set to items sharing the test item’s vocalized template and grammatical category, the remaining exemplars differ from one another only in consonant identity. Removing that distinction returns the model’s predictions to what a frequency count over the restricted set would yield.

It should be noted that neutralization corrects 199 predictions and introduces errors in 91, for the overall gain of 108 items that the 2.88% difference represents; in terms of plural types, 123 of corrections shift the prediction toward a sound plural and 76 toward a broken plural. Table 16 gives examples of each.

Singular	Full-feature GCM (incorrect)	Neutralized GCM (correct)
<i>fərfit^sa</i> ‘fork’	<i>CCaCəC</i>	sound - <i>at</i>
<i>mənzila</i> ‘status’	<i>CCaCəC</i>	sound - <i>at</i>
<i>bhima</i> ‘beast’	<i>CCaCəC</i>	sound - <i>at</i>
<i>dqiqa</i> ‘minute’	<i>CCaCəC</i>	sound - <i>at</i>
<i>bəlbuzə</i> ‘blister’	sound - <i>at</i>	<i>CCaCəC</i>
<i>bərt^sma</i> ‘flat’	sound - <i>at</i>	<i>CCaCəC</i>
<i>drəŋ</i> ‘arm’	<i>CCuC</i>	<i>CCuCa</i>
<i>bnin</i> ‘tasty’	sound - <i>in</i>	<i>CCaC</i>

Table 16: Examples of predictions corrected in GCM Restricted when consonantal features are neutralized. The upper block shows corrections toward a sound plural, the lower block corrections toward a broken plural.

The convergence of GCM Restricted (consonants neutralized) with the Vocalized Template match model at 69.13% reflects the fact that these two architectures implement the same underlying computation once consonantal identity is removed from the similarity metric. Both restrict plural selection to the vocalic-prosodic tier of the singular, discarding consonantal information entirely. This manipulation therefore establishes that the loss in accuracy incurred by the GCM model is due to including consonantal similarity in plural prediction.

5.4 Why Dawdy-Hesterberg and Pierrehumbert (2014) found a boost from segmental similarity

The finding that consonantal similarity provides no benefit—and reduces accuracy—in the present study seems to contradict Dawdy-Hesterberg and Pierrehumbert (2014), who found a small but significant accuracy boost when segmental similarity was added to their Simple Template Match model for MSA. A closer examination of the design differences between the two studies, however, suggests that the boost they observed is not evidence that segmental similarity is independently useful for generalization, but rather a consequence of their use of the abstract CV template as the restriction criterion.

The crucial design difference is that Dawdy-Hesterberg and Pierrehumbert (2014) restricted their GCM model using the abstract CV template, not the vocalized template. Because the abstract CV template does not specify vowel quality, the comparison set for a given test item contains training items that differ from one another—and from the test item—in vowel quality. In this setting, segmental similarity can be beneficial by prioritizing the weighting of training items that are vocally similar to the test item, effectively approximating what a vocalized template restriction achieves directly. The boost from segmental similarity in Dawdy-Hesterberg and Pierrehumbert (2014) is therefore better understood as the GCM recovering vocalic information that the abstract template restriction excluded, rather than as evidence that both vocalic and consonantal similarity are independently informative for generalization.

This hypothesis makes a direct prediction: when the GCM is restricted by the abstract CV template rather than the vocalized template, segmental similarity should help; but when it is restricted by the vocalized template, which already encodes vocalic information, segmental similarity should add nothing, or even cause harm due to consonantal noise. To test this, the GCM Restricted model was run twice, once with each restriction criterion. The two simulations are identical in every other respect: the same similarity metric, the same parameter settings, the same fold assignments, and the same corpus. The only difference is whether the comparison set is drawn from items sharing the test item’s abstract CV template or its vocalized template. Each simulation is then compared against the Template Match model that uses the same restriction, so that the effect of adding segmental similarity is measured separately at each level of abstraction.

A further manipulation isolates which component of the similarity metric is responsible for any effect. Following the neutralization procedure described above, each simulation was repeated twice: once with consonantal distinctions neutralized, so that only vocalic similarity can be registered, and once with vocalic distinctions neutralized, so that only consonantal similarity remains.

As shown in Table 17, the results are consistent with the hypothesis. When the restriction criterion is the abstract CV template, GCM Restricted outperforms the Template Match model, reproducing the pattern found by Dawdy-Hesterberg and Pierrehumbert (2014) for MSA, although the difference is small at (+0.81%) and does not itself reach significance ($\beta = 0.14$, $p = .145$). When the restriction criterion is the vocalized template, GCM Restricted underperforms the Template Match model, and this difference is significant ($\beta = -0.59$,

$p < .001$).

Restriction	Templ. Match	GCM Restricted		
		full	vowels only	cons. only
Abstract CV template	65.52%	66.33% (+0.81)	69.22% (+3.70)	65.69% (+0.17)
Vocalized template	69.13%	66.25% (-2.88)	69.13% (0.00)	66.27% (-2.86)

Table 17: Effect of segmental similarity on accuracy depending on the level of template abstraction used in the restriction. “Vowels only” neutralizes consonantal distinctions in the similarity metric; “consonants only” neutralizes vocalic distinctions. Differences from the corresponding Template Match model in parentheses.

Breaking down the similarity metric shows why, and makes the mechanism explicit. Under abstract restriction, a GCM whose consonantal distinctions have been neutralized, so that only vocalic similarity can be captured, reaches 69.22%, a gain of 3.70% over the Abstract Template match model, which is a significant improvement ($\beta = 1.11$, $p < .001$). This is actually the accuracy of the Vocalized Template match model itself (69.13%). The opposite manipulation shows the same pattern: a GCM whose vocalic distinctions have been neutralized, leaving only consonantal similarity, reaches 65.69%, against 65.52% for the Abstract Template match model, a difference of 0.17%.

The full-feature model’s +0.81 is therefore the sum of two opposing effects. With abstract templates, vocalic similarity recovers information the template leaves unspecified, but the consonantal component of the same metric contributes nothing and obscures that gain, driving the model from 69.22% back to 66.33%. With vocalized templates, the vowels are already fixed across the comparison set, so only consonantal similarity is left to influence predictions; neutralizing vowels there changes nothing (66.27% vs. 66.25%).

The reason the boost coming from segmental similarity in Dawdy-Hesterberg and Pierrehumbert (2014) is small rather than large is also consistent with this interpretation. Their GCM Restricted model does not neutralize consonantal features, meaning that consonantal similarity was computed together with vocalic similarity. As the neutralization test above shows, the small boost they observed reflects the result of two opposing effects: a positive contribution from vocalic similarity and a negative contribution from consonantal noise. If Dawdy-Hesterberg and Pierrehumbert (2014) had neutralized consonantal features and retained only vocalic similarity, the boost would likely have been larger, and the resulting model performance would have approximated what vocalized template restriction achieves directly.

Taken together, the findings of the two studies are compatible: segmental similarity appears to help in MSA only because the abstract CV template leaves vocalic information unspecified, and the GCM partially recovers it through similarity weighting. Once the template is vocalized, this approximation is no longer needed, and the only remaining component of segmental similarity (consonantal identity) introduces noise. It should also be noted that some of the difference in outcomes between the two studies may reflect methodological dif-

ferences. First, Dawdy-Hesterberg and Pierrehumbert (2014) used a 75%/25% training/test split iterated 10 times, with a minimum gang size of four items, while the present study uses standard 10-fold cross-validation with a 90%/10% split and no minimum gang size constraint. Second, they worked with unpointed Arabic orthography, in which short vowels are omitted, while the present study uses fully vocalized IPA transcriptions.⁶ Third, the two studies differ in the variety examined: MSA versus Moroccan Arabic, whose plural systems differ in complexity and fragmentation (Author, to appear).

It is worth adding that Dawdy-Hesterberg and Pierrehumbert (2014) identify the same gap this study addresses. They note that features of the vocalic melody of the singular appear to be correlated with the plural pattern, and that their ability to examine these factors was limited by their use of unvoweled text. The present results can be seen as addressing that observation using data in which the vocalic melody is fully specified.

5.5 The effect across the parameter range

Two facts about the sensitivity parameter s , introduced in Section 5.1, bear on how the results so far should be understood. The first is that the value used above, $s = 0.3$, is not a constant of the model but the value Dawdy-Hesterberg and Pierrehumbert (2014) found to fit their MSA data best; there is no guarantee that it is also the best value that would fit Moroccan Arabic. The second is that s governs how far the model departs from template matching: as s grows the similarity weights approach 1 for every exemplar and the gang score reduces to gang size, so that the GCM converges on the corresponding Template Match model, while as s decreases, weight concentrates on the nearest neighbors. Any difference between the two architectures is therefore a function of s , and neither the negative effect reported in Section 5.2 nor the contrast just described can be settled at a single value of it. Table 18 accordingly reports both restrictions across the range.

⁶Nieder et al. (2022) draw the same connection for Maltese: adding an abstract CV template to segmental information does not improve their classifiers' predictions of Maltese noun plural classes, and they attribute the contrast with Dawdy-Hesterberg and Pierrehumbert (2014)'s MSA result to precisely this orthographic difference, since Maltese orthography represents vowels fully.

s	Abstract restriction		Vocalized restriction	
	GCM	vs. TM	GCM	vs. TM
0.1	58.70%	−6.82	59.07%	−10.06
0.2	64.16%	−1.36	63.92%	−5.21
0.3	66.33%	+0.81	66.25%	−2.88
0.5	68.36%	+2.84	68.62%	−0.51
0.75	67.77%	+2.25	69.11%	−0.02
1.0	67.18%	+1.66	69.19%	+0.06
1.5	66.48%	+0.96	69.21%	+0.08
2.0	66.19%	+0.67	69.16%	+0.03
3.0	66.13%	+0.61	69.16%	+0.03
5.0	66.03%	+0.51	69.05%	−0.08
10.0	65.76%	+0.24	69.08%	−0.05

Table 18: GCM Restricted accuracy across values of the sensitivity parameter s , under each restriction criterion, with the difference from the corresponding Template Match model.

Three observations are worth noting. First, at every value of s tested, segmental similarity is worth more under abstract restriction than under vocalized restriction, by between 0.29% and 3.85%. The contrast reported above is therefore a property of the parameter space rather than of one setting. It should be added that at $s = 0.3$ the abstract-restriction benefit is not itself significant; the difference between the two restrictions is carried mainly by the negative effect under vocalized restriction, and the positive effect is largest at $s = 0.5$ rather than at the value fitted by Dawdy-Hesterberg and Pierrehumbert (2014) for MSA. Second, the benefit under abstract restriction is largest at small s and decays steadily toward zero as s grows, exactly as the interpretation predicts: the vocalic information that similarity weighting recovers is recovered by attending to the nearest neighbors, and a model that weights all exemplars equally cannot recover it. Third, under vocalized restriction the GCM never exceeds Template Match by more than 0.08% at any value of s . Segmental similarity provides no benefit there at any setting, and the negative effect at low s reflects the fact that the only variation left for the metric to weight is consonantal.

The parameter fitted to the present data would differ from 0.3 under both restrictions: the optimal value lies near 0.5 for the abstract restriction and 1.5 for the vocalized one. The value of Dawdy-Hesterberg and Pierrehumbert (2014) is retained in the tables above because the purpose of the comparison is to align the two studies, and a refitted parameter would confound the comparison with a methodological change. Nothing in the argument depends on the value chosen, as Table 18 shows.

6 Conclusion

This paper investigated the level of phonological abstraction at which morphological analogy operates in Moroccan Arabic. Four levels of representation were compared as bases for analogy: the consonantal root, the abstract CV template, the vocalized template, and the full segmental representation. The first three were implemented as a hierarchy of Template Match models evaluated by 10-fold cross-validation on a corpus of 3,735 singular–plural pairs. The vocalized template emerged as the best of the three, outperforming both the abstract CV template and the root. The gain falls almost entirely on the broken plurals. The complementary question—whether a less abstract representation incorporating fine-grained segmental similarity does better—did not receive support: a GCM restricted by the vocalized template never exceeded vocalized template matching at any setting of its sensitivity parameter, and fell substantially below it when similarity was weighted sharply. Neutralizing consonantal features returned it to the Template Match level exactly, since under vocalized restriction consonantal identity is the only distinction the metric can capture, and removing it leaves the model with nothing but gang size. Consonantal identity therefore contributes nothing to predicting plurals in Moroccan Arabic beyond what the vocalized template already encodes.

These results also resolve an apparent conflict with previous work on MSA. Dawdy-Hesterberg and Pierrehumbert (2014) found that segmental similarity improved prediction, which seems to contradict the present findings. Decomposing the similarity metric shows that its value depends on the restriction criterion. Under abstract restriction, a metric confined to vowels recovers almost all of the vocalic information the template leaves unspecified, while one confined to consonants leaves accuracy unchanged. Under vocalized restriction that information is already encoded, so nothing remains for the vocalic component to recover, only the consonantal component is left, and accuracy falls. The two sets of findings are compatible once this is recognized.

Taken together, these findings support a view of analogy in Arabic as operating over vocalized, prosodic template representations, with consonantal identity playing no measurable role in these models. This has a direct implication for analogical models of morphology more broadly. The GCM framework weights all phonological features equally in the similarity computation, which is appropriate when all features are informative but fails in Semitic morphology, where consonantal and vocalic features belong to independent morphological tiers with different functions (McCarthy, 1979, 1981): consonants encode lexical-semantic identity, while vowels and prosodic structure encode grammatical information including plural class membership. The success of the Vocalized Template match model points toward a model architecture in which the input is first abstracted to the vocalized template level before morphological analogy operates, ignoring consonantal identity while retaining vocalic and prosodic structure.

The accuracy of the Vocalized Template match model, 69.13%, is the best performance achieved by any model tested here. Whether a richer model could exceed it remains open; no attempt to do so was made in this study. What the present results do suggest is that the remaining errors are unlikely to be attributable to the choice of phonological representation,

and are more likely due to factors lying outside the template entirely: lexical idiosyncrasy, etymology, semantic class, and the within-word variation documented in Section 3.3.

This finding deviates from an extensive body of literature on lexical processing showing a consonant bias (C-bias): the finding that consonants carry more weight than vowels in word recognition and learning. This asymmetry has been documented across multiple languages and modalities, including auditory word recognition (Cutler et al., 2000; Bonatti et al., 2005), visual word recognition (New et al., 2008; New and Nazzi, 2014; Schmandt et al., 2022), and infant word learning (Nespor et al., 2003; Nazzi, 2005). The C-bias has been attributed to the greater lexical distinctiveness of consonants: across many languages, consonant neighborhoods are smaller than vowel neighborhoods, making consonants more informative for identifying individual words (Keidel et al., 2007; Nazzi and New, 2007). In the present study, however, the pattern is reversed: vocalic information is the primary driver of plural prediction, while consonantal identity contributes nothing that the vocalized template does not already encode.

This reversal is consistent with the functional organization of Semitic morphology, where consonantal roots encode lexical-semantic meaning and are shared across all inflectional forms of a word, while vocalic melodies and prosodic templates encode grammatical distinctions. The C-bias literature and the present findings are therefore not contradictory: they reflect the fact that whether consonants or vowels are informative depends on the linguistic task and the pattern under investigation. Consonants may be more informative for lexical identification, but vowels are more informative for morphological generalization in a system where grammatical information is encoded on the vocalic tier.

One qualification is important for interpreting these results. Ten-fold cross-validation measures how well a model predicts the plurals of held-out existing words. It is a measure of how much of the lexicon is recoverable from a given level of phonological representation, i.e., what information the learning data makes available to a learner. It is not, by itself, a measure of how speakers extend patterns to forms they have never encountered; that is an experimental question, and I leave it for future work.

References

- Al Ghadi, A. (1990). *Moroccan Arabic Plurals and the Organization of the Lexicon*. Faculty of Letters and Humanities, Mohammed V University.
- Albright, A. and Hayes, B. (2003). Rules vs. analogy in English past tenses: a computational/experimental study. *Cognition*, 90:119–161. [https://doi.org/10.1016/S0010-0277\(03\)00146-X](https://doi.org/10.1016/S0010-0277(03)00146-X).
- Alegre, M. and Gordon, P. (1999). Rule-based versus associative processes in derivational morphology. *Brain and language*, 68(1-2):347–354. <https://doi.org/10.1006/brln.1999.2066>.
- Alrashed, F. (2021). *Investigation of the Distributions, Derivation, and Generalizations in*

- Arabic Plural System*. PhD thesis, University of Michigan. <https://doi.org/10.7302/1465>.
- Baayen, H. (1992). Statistical models for word frequency distributions: A linguistic evaluation. *Computers and the Humanities*, 26(5):347–363. <https://doi.org/10.1007/BF00136980>.
- Becker, M. and Gouskova, M. (2016). Source-oriented generalizations as grammar inference in russian vowel deletion. *Linguistic inquiry*, 47(3):391–425. https://doi.org/10.1162/LING_a_00217.
- Becker, M., Ketrez, N., and Nevins, A. (2011). The surfeit of the stimulus: Analytic biases filter lexical statistics in turkish laryngeal alternations. *Language*, 87(1):84–125. <https://doi.org/10.1353/lan.2011.0016>.
- Benhallam, A. (1980). *Syllable Structure and Rule Types in Arabic*. PhD thesis, University of Florida.
- Bensoukas, K. and Boudlal, A. (2012). The prosody of moroccan amazigh and moroccan arabic: Similarities in the phonology of schwa. *Prosody matters: essays in honor of Lisa Selkirk*, pages 3–42. <https://doi.org/10.3138/9781845536770>.
- Berko, J. (1958). The child’s learning of English morphology. *Word*, 14:150–177. <https://doi.org/10.1080/00437956.1958.11659661>.
- Bonatti, L. L., Peña, M., Nespor, M., and Mehler, J. (2005). Linguistic constraints on statistical computations: The role of consonants and vowels in continuous speech processing. *Psychological Science*, 16(6):451–459. <https://doi.org/10.1111/j.0956-7976.2005.01556.x>.
- Boudelaa, S. and Gaskell, M. G. (2002). A re-examination of the default system for arabic plurals. *Language and cognitive processes*, 17(3):321–343. <https://doi.org/10.1080/01690960143000245>.
- Boudelaa, S. and Marslen-Wilson, W. D. (2004). Abstract morphemes and lexical representation: the cv-skeleton in arabic. *Cognition*, 92(3):271–303. <https://doi.org/10.1016/j.cognition.2003.08.003>.
- Boudlal, A. (2001). *The Prosody and Morphology of a Moroccan Arabic Dialect: An Optimality-Theoretic Account*. VDM Verlag, Rabat.
- Bounhar, A. and Majjodi, A. E. (2025). Atlaset dataset for moroccan darija: From data collection, analysis, to model trainings. *Hugging Face Blog*. <https://huggingface.co/blog/atlasia/atlasat-dataset-moroccan-darija>.
- Bybee, J. (2001). *Phonology and language use*. Cambridge University Press. <https://doi.org/10.1017/CB09780511612886>.

- Cutler, A., Sebastián-Gallés, N., Soler-Vilageliu, O., and Van Ooijen, B. (2000). Constraints of vowels and consonants on lexical selection: Cross-linguistic comparisons. *Memory & Cognition*, 28(5):746–755. <https://doi.org/10.3758/BF03198409>.
- Daelemans, W., Gillis, S., and Durieux, G. (1994). The acquisition of stress: A data-oriented approach. *Computational Linguistics*, 20(3):421–453. <https://aclanthology.org/J94-3007/>.
- Dawdy-Hesterberg, L. G. (2015). *The structural and statistical basis of morphological generalization in Arabic*. PhD thesis, Northwestern University.
- Dawdy-Hesterberg, L. G. and Pierrehumbert, J. B. (2014). Learnability and generalisation of arabic broken plural nouns. *Language, cognition and neuroscience*, 29(10):1268–1282. <https://doi.org/10.1080/23273798.2014.899377>.
- Ernestus, M. and Baayen, R. H. (2003). Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language*, 79(1):5–38. <https://doi.org/10.1353/lan.2003.0076>.
- Ferguson, C. A. (1959). Diglossia. *word*, 15(2):325–340. <https://doi.org/10.1080/00437956.1959.11659702>.
- Frisch, S., Pierrehumbert, J., and Broe, M. (2004). Similarity avoidance and the OCP. *Natural Language and Linguistic Theory*, 22:179–228. <https://doi.org/10.1023/B:NALA.00000005557.78535.3c>.
- Haeri, N. (2003). *Sacred language, ordinary people: Dilemmas of culture and politics in Egypt*. Springer. <https://doi.org/10.1057/9780230107373>.
- Hammond, M. (1988). Templatic transfer in Arabic broken plurals. *Natural Language and Linguistic Theory*, 6(2):247–270. <https://doi.org/10.1007/BF00134231>.
- Harrell, R. (1962). *A Short Reference Grammar of Moroccan Arabic*. Georgetown University Press.
- Harrell, R. and Sobelman, H. (1966). *A Dictionary of Moroccan Arabic: Arabic-English*. Georgetown University Press, Washington, D.C.
- Heath, J. (1987). *Ablaut and Ambiguity: Phonology of a Moroccan Arabic dialect*. State University of New York Press, Albany. <https://doi.org/10.2307/414937>.
- Idrissi, A., Alazbi, S., and Marzouki, Y. (2025). Tracking the stem and root morphemes in arabic: Evidence from visual morphological priming. *Journal of Psycholinguistic Research*, 54(6):1–29. <https://doi.org/10.1007/s10936-025-10173-1>.
- Jarosz, G., Hughes, C., Lamont, A., Prickett, B., Baird, M., Kim, S., and Nelson, M. (2025). Type and token frequency jointly drive learning of morphology. *Journal of Memory and Language*, 144:104666. <https://doi.org/10.1016/j.jml.2025.104666>.

- Keidel, J., Jenison, R., Kluender, K., and Seidenberg, M. (2007). Does grammar constrain statistical learning? commentary on Bonatti, Peña, Nespore, and Mehler (2005). *Psychological Science*, 18(10):922–923. <https://doi.org/10.1111/j.1467-9280.2007.02001.x>.
- Keuleers, E., Sandra, D., Daelemans, W., Gillis, S., Durieux, G., and Martens, E. (2007). Dutch plural inflection: The exception that proves the analogy. *Cognitive psychology*, 54(4):283–318. <https://doi.org/10.1016/j.cogpsych.2006.07.002>.
- Kruskal, J. B. (1983). An overview of sequence comparison: Time warps, string edits, and macromolecules. *SIAM review*, 25(2):201–237. <https://doi.org/10.1137/1025045>.
- Linzen, T., Kasyanenko, S., and Gouskova, M. (2013). Lexical and phonological variation in russian prepositions. *Phonology*, 30(3):453–515. <https://doi.org/10.1017/S0952675713000225>.
- Marcus, G. F., Brinkmann, U., Clahsen, H., Wiese, R., and Pinker, S. (1995). German inflection: The exception that proves the rule. *Cognitive psychology*, 29(3):189–256. <https://doi.org/10.1006/cogp.1995.1015>.
- Mashaqba, B., Al-Khawaldeh, N., AlGweirien, H., and Al-Edwan, Y. (2020). Acquisition of broken plural patterns by jordanian children. *Linguistics*, 58(4):1009–1022. <https://doi.org/10.1515/ling-2020-0024>.
- McCarthy, J. J. (1979). On stress and syllabification. *Linguistic Inquiry*, 10(3):443–465.
- McCarthy, J. J. (1981). A Prosodic Theory of Nonconcatenative Morphology. *Linguistic inquiry*, 12(3).
- McCarthy, J. J. and Prince, A. S. (1990). Foot and Word in Prosodic Morphology: The Arabic Broken Plural. *Natural Language and Linguistic Theory*, 8(2):209–283. <https://doi.org/10.1007/BF00208524>.
- McCurdy, K. (2024). *Rules, frequency, and predictability in morphological generalization: behavioral and computational evidence from the German plural system*. PhD thesis, The University of Edinburgh. <https://doi.org/10.7488/era/4161>.
- Nakisa, R. C., Plunkett, K., and Hahn, U. (1998). A cross-linguistic comparison of single and dual-route models of inflectional morphology. *cognitive models of language acquisition*, ed. by p. broeder, and j. murre.
- Nakisa, R. C., Plunkett, K., and Hahn, U. (2000). Single and dual-route models of inflectional morphology. <https://doi.org/10.1093/oso/9780198299899.003.0010>.
- Nazzi, T. (2005). Use of phonetic specificity during the acquisition of new words: differences between consonants and vowels. *Cognition*, 98(1):13–30. <https://doi.org/10.1016/j.cognition.2004.10.005>.
- Nazzi, T. and New, B. (2007). Beyond stop consonants: Consonantal specificity in early lexical acquisition. *Cognitive Development*, 22(2):271–279. <https://doi.org/10.1016/j.cogdev.2006.10.007>.

- Nespor, M., Peña, M., and Mehler, J. (2003). On the different roles of vowels and consonants in speech processing and language acquisition. *Lingue e linguaggio, Rivista semestrale*, (2/2003):203–230. <https://doi.org/10.1418/10879>.
- New, B., Araújo, V., and Nazzi, T. (2008). Differential processing of consonants and vowels in lexical access through reading. *Psychological Science*, 19(12):1223–1227. <https://doi.org/10.1111/j.1467-9280.2008.02228.x>.
- New, B. and Nazzi, T. (2014). The time course of consonant and vowel processing during word recognition. *Language, Cognition and Neuroscience*, 29(2):147–157. <https://doi.org/10.1080/01690965.2012.735678>.
- Nieder, J., Tomaschek, F., Cohrs, E., and de Vijver, R. v. (2022). Modelling maltese noun plural classes without morphemes. *Language, Cognition and Neuroscience*, 37(3):381–402. <https://doi.org/10.1080/23273798.2021.1977835>.
- Nieder, J., van de Vijver, R., and Mitterer, H. (2021). Knowledge of maltese singular–plural mappings: Analogy explains it best. *Morphology*, 31(2):147–170. <https://doi.org/10.1007/s11525-020-09353-7>.
- Nirheche, A. (2021). Broken plural formation in Moroccan Arabic: a stratal optimality theory account. Master’s thesis, Memorial University of Newfoundland.
- Nirheche, A. (2025). Variable and exceptional assimilation of the definite article [l-] in moroccan arabic. In *Proceedings of the Annual Meetings on Phonology*, volume 1. University of Massachusetts Amherst Libraries.
- Nirheche, A. and Becker, M. (2025). Nonfinality drives epenthesis in the moroccan arabic broken plural. *Brill’s Journal of Afroasiatic Languages and Linguistics*, 17(2):333–359.
- Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, 115(1):39. <https://doi.org/10.1037/0096-3445.115.1.39>.
- Outchakoucht, A. and Es-Samaali, H. (2021). Moroccan dialect -darija- open dataset. <https://github.com/darija-open-dataset/dataset>.
- Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In Bybee, J. and Hopper, P., editors, *Frequency and the Emergence of Linguistic Structure*, pages 135–157. John Benjamins.
- Plunkett, K. and Nakisa, R. C. (1997). A connectionist model of the arabic plural system. *Language and Cognitive processes*, 12(5-6):807–836. <https://doi.org/10.1080/016909697386691>.
- Ratcliffe, R. R. (2002). The broken plural system of moroccan arabic: Diachronic and cognitive perspectives. In Parkinson, D. B. and Benmamoun, E., editors, *Perspectives on Arabic Linguistics: Papers from the Annual Symposium on Arabic Linguistics. Volume XIII-XIV: Stanford, 1999 and Berkeley, California 2000*, pages 87–109. John Benjamins Publishing Company. <https://doi.org/10.1075/cilt.230.06rat>.

- Rácz, P., Beckner, C., Hay, J. B., and Pierrehumbert, J. B. (2020). Morphological convergence as on-line lexical analogy. *Language*, 96(4):735–770. <https://doi.org/10.1353/lan.2020.0061>.
- Sabri, A.-S., Huneety, A., Mashaqba, B., and Zuraiq, W. (2020). Similarity effects on the emergence of default inflection in jordanian arabic. *Jordan Journal of Modern Languages and Literatures Vol*, 12(3):291–305.
- Schmandt, S., Nazzi, T., and New, B. (2022). Consonant, vowel and lexical neighbourhood processing during word recognition: New evidence using the sandwich priming technique. *Language, Cognition and Neuroscience*, 37(9):1115–1130.
- Skousen, R. (1989). *Analogical Modeling of Language*. Dordrecht: Kluwer. <https://doi.org/10.1007/978-94-009-1906-8>.