Variable and Exceptional Assimilation of the Definite Article [l-] in Moroccan Arabic*

Ali Nirheche
University of Massachusetts Amherst

1 Introduction

The investigation of variable and exceptional phonological patterns has become an area of growing interest in phonological theory. Research has focused on the representation and learning of phenomena involving both variable and exceptional patterns (Hayes & Londe, 2006; Pater et al., 2012; Linzen et al., 2013; Shih, 2018; Hughto et al., 2019). This paper contributes to this research area by investigating the morpheme-specific assimilation patterns of the Moroccan Arabic definite article [l-], specifically focusing on the variable assimilation observed in [ʒ]-initial words.

As in Modern Standard Arabic (MSA) and other Arabic varieties, the definite article [l-] in Moroccan Arabic exhibits morpheme-specific behavior, in the form of total assimilation (gemination), when attached to a coronal-initial word (Harrell, 1962; Heath, 1987, 1989; Maas & Procházka, 2022), as in [l-dar] → [ddar] 'house'. Uniquely, in Moroccan Arabic, assimilation does not uniformly occur with the palatal fricative [ʒ], leading to variation in assimilation with some [ʒ]-initial words triggering it (e.g. [ʒʒar] 'DEF-neighbor'), while other words resisting it (e.g. [lʒil] 'DEF-generation'). A corpus study I conducted reveals that the assimilation of [ʒ]-initial words is phonologically-conditioned by the sound following [ʒ], with words where a consonant follows [ʒ] are more likely to assimilate than those where a schwa follows [ʒ], which, in turn, are more likely to assimilate than those where a vowel follows [ʒ].

A critical question that this paper addresses is how Moroccan Arabic speakers generalize the assimilation patterns to [ʒ]-initial nonce words. Through a nonce word experiment, I show that the assimilation patterns for nonce words align with the distributions observed across the lexicon. Previous studies on the productivity of morphophonological alternations have shown that, when speakers generalize to nonce forms, they tend to follow the lexical statistics (Zuraw, 2000; Ernestus & Baayen, 2003; Hayes & Londe, 2006; Hayes, 2009; Linzen et al., 2013; Becker & Gouskova, 2016). The experimental results presented in this paper align with these studies.

One challenge that learning theories face is accounting for both the stochastic behavior of nonce words and the fixed pronunciations of real words. While some previous studies, such as Hayes & Wilson (2008), offered solutions that were not applicable to alternations, other studies offered more promising solutions to this problem (Becker & Gouskova, 2016; Moore-Cantwell & Pater, 2016; Hughto et al., 2019). The definite article assimilation patterns in Moroccan Arabic provide an ideal test case for learning models given the complex exceptionality and variation patterns it embodies. In this paper, I implement a Maximum Entropy grammar (Goldwater & Johnson, 2003:MaxEnt) together with lexically-indexed constraints (Pater, 2000, 2009), which accounts for exceptionality by the use of constraints indexed to exceptional lexical items. I will show that the proposed model can successfully predict variation for [ʒ]-initial nonce words as well as a fixed categorical behavior for real words. I will also show that the model predicts that both [ʒ]-initial words that assimilate and those that do not have some degree of exceptionality.

* Many thanks to Michael Becker and Gaja Jarosz for their invaluable discussions and insights. Thanks also to Joe Pater, Karim Bensoukas, and Ali Idrissi for their helpful comments. I'd also like to thank the UMass Sound Workshop and the audience at the 2024 Annual Meeting on Phonology for their feedback. All errors remain my own.

Proceedings of *AMP 2023/2024*

The rest of the paper is organized as follows: Section 2 illustrates the patterns observed within definite article assimilation in Moroccan Arabic with a particular focus on the behavior of [ʒ]-initial words, presenting a corpus study to identify factors influencing assimilation for these words. Section 3 proposes a MaxEnt learning model with lexically-indexed constraints that accurately predicts both the variation predicted for [ʒ]-initial nonce words and fixed behavior of real words. Section 4 presents the results of a nonce word experiment that support frequency marching, indicating that the assimilation of [ʒ]-initial nonce words varies and aligns with lexical regularities. Section 6 concludes.

2 The Exceptional and Variable Patterns of Definite Article Assimilation

2.1 *Definite article assimilation and ʒ-initial words* In Moroccan Arabic, definite nouns are formed by attaching the prefix [l-] to a given noun (1). When the noun begins with a CV sequence where C is a non-coronal consonant, the definite article is attached faithfully (1a). When the noun begins with a CC sequence where the initial C is a non-coronal consonant, a schwa is inserted between the definite article and the noun (1b). When the noun begins with a coronal consonant (either in a CC or CV sequence), however, attaching [l-] results in the total assimilation of the definite article to the initial coronal consonant, forming a geminate sound[1] (Harrell, 1962; Heath, 1987, 1989; Maas & Procházka, 2022) as seen in (1c).

(1)

	Noun	**Definite Noun**	**Gloss**
a.	bənt	l.bən.t	‘girl’
	kora	l.ko.ra	‘ball’
b.	fraʃ	ləf.raʃ	‘bed sheets’
	kbal	lək.bal	‘popcorn’
c.	dar	ddar	‘house’
	nhar	nn.har	‘day’
	ʃəmʃ	ʃʃəm.ʃ	‘sun’
	lil	llil	‘night’

The assimilation pattern seen in (1c) is morpheme specific behavior, since it is not triggered word-internally or in the presence of other morphemes like the preposition prefix [l-] that attaches to nouns to form prepositional phrases (2).

(2)

	/l/+/dar/+/o/	ldaro	*ddaro	‘to-house-POSS.3SG’
	/l/+/nhar/	lənhar ssəbt	*nnhar ssəbt	‘to-day Saturday’

Interestingly, assimilation of the definite article [l-] to nouns beginning with [ʒ] is variable (Heath, 1987, 1989; Maas & Procházka, 2022). The palatal fricative [ʒ] triggers assimilation in some words (3a) and does not in others (3b). This is an unusual exception given the fact that [ʃ], which only differs from [ʒ] in voicing, categorically triggers assimilation.

(3)

	Noun	**Definite Noun**	**Gloss**
a.	ʒoqa	ʒʒoqa	‘gathering’
	ʒməl	ʒʒməl	‘camel’
	ʒuʕ	ʒʒuʕ	‘hunger’
b.	ʒomhor	lʒomhor	‘audience’
	ʒəlsa	lʒəlsa	‘court session’

Previous work on definite article assimilation in Moroccan Arabic classifies cases like (3b) as exceptions to the categorical assimilation rule. According to Harrell (1962), [ʒ]-initial words that do not trigger assimilation are those that belong to religious terminology. Freeman (2016:p. 177) claims that the failure of [ʒ] to trigger assimilation in some Moroccan Arabic words is “the result of diglossic interference from

[1] I assume that geminites are single sounds. The same assumption has been made for Tashlhiyt Berber, spoken in Morocco, where a geminite sound belongs to the onset of the same syllable ([llan] ‘they exist’), whereas non-identical CC sequences are parsed into two separate syllables ([g.ru] ‘glean’) (Ridouane, 2016).

the standard or classical language". Heath (1987) argues that assimilation with [ʒ]-initial words is default for inherent Moroccan Arabic words, while all MSA loans beginning with [ʒ] do not trigger assimilation. There are two main issues with these previous proposals. First, there are words that are borrowed from other languages or that do not belong to religious/political register, but still fail to trigger assimilation (4a). Second, the issue with Heath's proposal is identifying what these "MSA loans" are, which is by no means a straightforward task. In fact, even what one might call "inherent" Moroccan Arabic words are historically derived from Classical Arabic. What's more, MSA words are being borrowed continously, so it's not an easy task to specify how old a word should be in order to be an inherent Moroccan Arabic word. It can be seen in (4b) that some words that do not seem to be recent borrowings from MSA do not assimilate. It is also unclear how the learner would identify these borrowed words to treat them exceptionally.

(4)

	Noun	**Definite Noun**	**Gloss**
a.	ʒø	l-ʒø	'game'
	ʒo ɣ	l-ʒo ɣ	'genre'
b.	ʒuw	ʒʒuw	'weather'
	ʒud	l-ʒud	'generosity'

2.2 *Assimilation of [ʒ]-initial words is phonologically conditioned* A possibility that was not discussed in the literature is for assimilation to be phonologically-conditioned. One factor that is worth examining is the following context, i.e. the sound following [ʒ]. There are three relevant categories of sounds: consonants, full vowels, and the schwa. The expected behavior of [ʒ]-initial words is for the words with a consonant following [ʒ] to assimilate more than the ones with a schwa following [ʒ] which are, in turn, expected to assimilate more than the words with a full vowel following [ʒ]. The reason that assimilation is more likely to occur when a consonant follows [ʒ] is the restriction against having three adjacent consonants. The definite form of the noun [ʒmil] is [ʒʒmil] because the non-assimilated version [lʒmil] begins with a sequence of three adjacent consonants. When a vowel follows [ʒ], however, assimilation is less likely to occur given the absence of a CCC sequence. Such sequence is avoided in non-coronal initial contexts through schwa epenthesis. The non-assimilated definite form version of [ʒil], which is [lʒil], begins with a sequence of two consonants, which is acceptable in Moroccan Arabic. When a schwa follows [ʒ], assimilation may have an intermediate rate of assimilation given the ambiguous status of the schwa in Moroccan Arabic (Benhallam, 1980; Al Ghadi, 1990; Boudlal, 2001; among others). The Moroccan Arabic schwa is often argued to be an epenthetic vowel only inserted for syllabification purposes, and whose status in the phonological grammar is different from that of a full vowel.

In order to investigate this factor, a corpus of 120 Moroccan Arabic [ʒ]-initial words was created and examined. The corpus is representative of the Moroccan Speaker's knowledge of [ʒ]-initial words given the various sources used for data collection. These sources include the author's knowledge as a native speaker of Moroccan Arabic, consulting other native speakers, a large online corpus of Moroccan Arabic: Darija Open Dataset (Outchakoucht & Es-Samaali, 2021), previous work that discussed this phenomenon (Harrell, 1962; Freeman, 2016; Maas & Procházka, 2022), and a Moroccan Arabic-English dictionary (Harrell & Sobelman, 1966). Overall, 76 out of the 120 words in the corpus exhibit assimilation. It can be seen in Table 1 that, as was predicted, assimilation is more likely to occur when a consonant follows [ʒ], is less likely to occur when a full vowel follows [ʒ], and has an intermediate rate of assimilation when a schwa follows [ʒ].

Following Context	**Count**	**Assimilation (%)**
Consonant	26	96%
Schwa	38	81%
Full Vowel	57	37%
total:	120	63%

Table 1: Corpus Statistics about the Assimilation Patterns based on the following context

The results in Table 1 were supported by the logistic regression analysis with custom contrasts, conducted using the `lme4` package (Bates et al., 2015) in R (R Development Core Team, 2014), to

determine if the following context significantly influences assimilation of [ʒ]-initial words. The dependent variable was the binary response (assimilation vs. no assimilation), and the independent variable was the `FollowingSound`, which was coded using Helmert contrasts. Specifically, the contrasts compared the schwa with consonant(*Schwa vs. Consonant*) and the vowel with the average of schwa and consonant contexts (*Vowel vs. Average of Schwa and Consonant*). The analysis revealed a statistically significant effect for the contrast "Vowel vs. Average of Schwa and Consonant" ($\beta = -2.49$, $SE = 0.5$, $t = -4.94$, $p < 0.00001$), which indicates a strong negative effect of a full vowel on assimilation. The contrast "Schwa vs. Consonant" was not statistically significant ($\beta = -0.9968$, $SE = 0.84$, $t = -1.17$, $p = 0.23$).

The next section presents a MaxEnt learning model that predicts the variability observed in the lexicon for [ʒ]-initial nonce words and accounts for the exceptionality patterns observed in known [ʒ]-initial words using lexically-indexed constraints.

3 Learning the Assimilation Patterns with a MaxEnt Model

3.1 *MaxEnt with lexically-indexed constraints* It has been shown that assimilation in definite nouns is a morpheme-specific phenomenon. It has also been shown that [ʒ]-initial words vary, with some words triggering assimilation and others resisting it. To account for this exceptionality and variation, I use MaxEnt grammar (Goldwater & Johnson, 2003) together with lexically-indexed constraints (Pater, 2000, 2009). MaxEnt is a probabilistic model that captures categorical and variable patterns in phonology. It assigns probabilities to different output candidates based on weighted constraints. Lexically-indexed constraints will also be proposed. Lexical indexation explains exceptionality by allowing constraints to be lexically specific, i.e. they apply only to certain lexical items or morphemes, not across the grammar. In our case, general constraints will account for the default behavior of the sequence of [l] followed by a coronal, where assimilation does not occur, and lexically-indexed constraints will account for the exceptionality of the definite article and the variable behavior of [ʒ]-initial words.

Let's consider an example of how this grammar works. We have seen in section 2 that, in the absence of the definite article, the sequence of [l] followed by a coronal does not trigger assimilation. To achieve this outcome, I propose an analysis in which the faithfulness constraint MAX(lat) interacts with the markedness constraint *l[cor]. MAX(lat), which requires identity of the lateral feature between the input and output, must have a higher weight compared to *l[cor], which penalizes outputs with the sequence of [l] followed by a coronal, in order to prevent assimilation. Three possible candidates are relevant: a non-assimilated, assimilated, and epenthesizing candidates. The latter, which epenthesizes a schwa between [l] and the following coronal, always loses when the noun following [l] begins with a simple onset. Therefore, the constraint *ə]$_\sigma$, which is never violated in Moroccan Arabic, is proposed to rule out the epenthesizing candidate in this context. MAX(lat), *l[cor] and *ə]$_\sigma$ are defined in (5).

(5)

- **MAX(lat)**: Assign a violation mark for every removal of the feature [lateral] from the output form.
- ***l[cor]**: Assign a violation mark for every lateral approximant followed by a coronal consonant in the output.
- ***ə]$_\sigma$**: Assign a violation mark for every schwa that surfaces in an open syllable.

An analysis in which *l[cor] has a lower weight than both MAX(lat) and *ə]$_\sigma$ predicts that assimilation cannot occur by default to any sequence of [l] followed by a coronal consonant. In order to account for the assimilation that is exceptionally triggered in the presence of the definite article [l], I propose a lexically-indexed version of *l[cor]:

(6)

- ***l[cor]$_{\text{DEF}}$**: Assign a violation mark to any instance of a lateral approximant followed by a coronal consonant that contains a phonological exponent of a morpheme specified as DEF.

By assigning a high weight to *l[cor]$_{\text{DEF}}$, assimilation is predicted to occur in definite nouns. The tableaux in (7) show the derivation of the prepositional phrase [ldaro] "to his house" and the definite noun [ddar] "the house" using this grammar.

(7)

		*l[cor]$_{\text{DEF}}$ $w=13$	*l[cor] $w=1$	MAX(lat) $w=5$	DEP $w=1$	*ə]$_\sigma$ $w=12$	$\mathscr{H}$	p
/l/ + /daro/	ddaro	0	0	−1	0	0	−5	≈1
	ldaro	0	−1	0	0	0	−1	≈0
	lədaro	0	0	0	−1	−1	−13	≈0
/l$_{\text{DEF}}$/ + /dar/	ddar	0	0	−1	0	0	−5	≈1
	ldar	−1	−1	0	0	0	−14	≈0
	lədar	0	0	0	−1	−1	−13	≈0

In (7), when [l-] is the preposition, the fully faithful candidate [ldaro] wins by satisfying the highly weighted faithfulness constraint MAX(lat). [ddaro], on the other hand, violates MAX(lat) by deleting the lateral feature of the input consonant /l/. Since MAX(lat) has a higher weight than *l[cor], the non-assimilating candidate [ldaro] is more harmonic than the assimilating one. Candidate [lədaro], which epenthesizes the schwa between the preposition [l] and coronal consonant [d], satisfies *l[cor], but it violates the highly weighted constraint *ə]$_\sigma$, and is, thus, ruled out. When [l-] is the definite article, however, the assimilated candidate (7a) only violates MAX(lat), resulting in a harmony score of -5. The fully faithful candidate, on the other hand, loses by violating *l[cor]$_{\text{DEF}}$ which has a high weight. The epenthesizing candidate also lose, since it violates the highly weighted constraint *ə]$_\sigma$. As can be seen, adding a lexically-indexed version of *l[cor] results in the morpheme-specific assimilation observed when forming definite nouns.

3.2 *The learning model* We have seen in section 2.2 that statistics across the lexicon suggest that assimilation is variable for [ʒ]-initial words based on the following context. Previous studies on the productivity of morphophonological alternations have shown that, when speakers generalize to nonce forms, they tend to follow lexical statistics, i.e. frequency match (Zuraw, 2000; Ernestus & Baayen, 2003; Hayes & Londe, 2006; Hayes, 2009; Linzen et al., 2013; Becker & Gouskova, 2016). In this section, I implement a MaxEnt model that will be shown to predict frequency matching behavior for nonce words. The model learns the grammar by exposure to the data without providing any information about what is default and what is exceptional with respect to the behavior of [ʒ]-initial words. I will show that the proposed MaxEnt model generates weights for a set of general and lexically-indexed constraints that predict the lexical trends observed for nonce words as well as the fixed categorical behavior of the existing [ʒ]-initial words. In terms of exceptionality, it will be shown that the model predicts that both [ʒ]-initial words that assimilate and those that do not have some degree of exceptionality.

To examine the learnability of the variable patterns of assimilation of [ʒ]-initial words, I used a MaxEnt learning model incorporating lexically-indexed constraints. The MaxEnt implementation that was used is Harmonic Grammar in R (Staubs, 2011:HGR), an algorithm that was created to run computations in Harmonic Grammar (Legendre et al., 1990a,b; Legendre & Smolensky, 2006; Boersma & Pater, 2016) using R (R Development Core Team, 2014). HGR uses a batch optimization algorithm that is guaranteed to converge on both probabilistic and categorical distributions as long as the model has all of the information relevant to the examined pattern. I used the L-BFGS-B optimization algorithm in conjunction with L2 regularization to find the constraint weights.

3.3 *Training data* The training data consists of 104 unique items: 25 items with the preposition [l-] attached to stems beginning with all possible conditions, 49 items with the definite article [l-] attached to non-[ʒ]-initial stems, and 34 items with the definite article [l-] attached to [ʒ]-initial words. Each input has three different outputs: a fully faithful candidate, an assimilating candidate, and an epenthesizing candidate. The proportion of items with each following context as well as their outcome (faithfulness, assimilation or epenthesis) matched the proportions in the lexical statistics.

3.4 *Constraints* In addition to MAX(lat), *l[cor] and *ə]$_\sigma$, the following general constraints were provided to the learning model: *#ʒʒ, DEP, *CCC, IDENT(cor), and *#CCə. The learning model was also given the following constraints that are lexically-indexed to the definite article [l]: *l[cor]$_{\text{DEF}}$ and DEP$_{\text{DEF}}$. The latter is needed to account for the behavior of the words that begin with a CC sequence. While the epenthesizing forms are optimal in words beginning with a CC sequence in the default context, as seen in

(1b) and (2), when the definite article is present, assimilation, not schwa epenthesis, occurs in such examples. This requires the use of an indexed version of DEP that prevents epenthesis specifically in the context of definite nouns.

(8)

- ***CCC**: Assign a violation mark for any sequence of three adjacent consonants in the output form.
- ***#ʒʒ**: Assign a violation mark for any output that begins with a gemintate [ʒ].
- **IDENT(cor)**: Assign a violation mark to any output sound whose corresponding input sound has a different value for the feature [coronal].
- ***#CCə**: Assign a violation mark to any output that begins with the sequence CCə.
- ***l[cor]$_{\text{DEF}}$**: Assign a violation mark to any instance of a lateral approximant followed by a coronal consonant that contains a phonological exponent of a morpheme specified as DEF.
- **DEP$_{\text{DEF}}$**: Assign a violation mark to any inserted segment that is adjacent to the morpheme specified as DEF[2].

While the indexation of the definite article was manual, the indexations for the nouns in the training data were automatic. The model was given versions of *l[cor], MAX(lat) and DEP that are lexically-indexed to all 104 input forms. In other words, each input had a lexically-indexed version of each of the three constraints (e.g. *l[cor]$_{\text{ʒàr}}$, MAX(lat)$_{\text{ʒàr}}$, and DEP$_{\text{ʒàr}}$ for the input */ʒar/*). This step is necessary to make the model determine which [ʒ]-words to treat as default and which as exceptional in an unsupervised manner.

3.5 *Results for real words* The model learned a grammar that accounts for the assimilation patterns of the sequence [l] followed by a coronal in all possible contexts. It learned the necessary weights for both the general and lexically-indexed constraints. The general constraints accounted for all cases involving the preposition [l] as well as the definite article when attached to items that begin with non-coronals or non-[ʒ] coronals. The model did assign weight to the indexed constraints of some non-[ʒ]-initial words that range between 1 and 5.3, but, interestingly, these weights were not necessary since the categorical behavior of these items was predicted using the general constraints only. Table 2 shows the weights of general constraints as well as the manually indexed constraints *l[cor]$_{\text{DEF}}$ and DEP$_{\text{DEF}}$.

*CCC	IDENT(cor)	*l[cor]$_{\text{DEF}}$	*ə]$_{\sigma}$	DEP	DEP$_{\text{DEF}}$	*ʒʒ	MAX(lat)	*#CCə	*l[cor]
24	15.4	12.5	11.2	8.7	6.7	5.9	5.9	0.9	0.02

Table 2: Generated weights of general and manually indexed constraints

With respect to existing [ʒ]-initial items, as shown in Table 3, the model generated weights for indexed constraints of both assimilated and non-assimilated [ʒ]-initial words. The results show a three way distinction based on the following context. In a vowel following context, for instance, the assimilated forms a weight of 7.3 for the indexed versions of *l[cor]. On the other hand, in a consonant following context, the non-assimilated forms had a weight of 10.8 for the indexed versions of MAX(lat). Therefore, the model treated both assimilated and non-assimilated [ʒ]-initial words as exceptional, although the decision of which constraint to indexed to which item and the weight value varied across the three contexts. This outcome seems to be required in order to get the fixed behavior of known [ʒ]-initial items.

[2] If we assume the traditional definitions of indexed constraints proposed by Pater (2000, 2009), DEP$_{\text{DEF}}$ would raise a locality problem since it's not directly associated with an exponent of the definite article, but with a segment adjacent to it (the schwa). Therefore, DEP$_{\text{DEF}}$ is defined in a way that specifically reflect this difference.

		Weights		
Context	**Assimilates**	*l[cor]	Max(lat)	Dep
ʒV	Yes	7.3	0	1.1
	No	0	8.1	1.1
ʒə	Yes	5.8	0	1.2
	No	0	9.2	1.2
ʒC	Yes	1.2	0	4.8
	No	1.2	10.8	0

Table 3: Weights of indexed-constraints based on the sound following ʒ and assimilation status

3.6 *Results for nonce words* In this paper, I follow Moore-Cantwell & Pater (2016) in their assumption about nonce words, i.e. that they would not be lexically-indexed. Therefore, the predictions for nonce words can be automatically determined, since the lexically-indexed constraints do not have an effect on nonce words. The model predictions show that both assimilated and non-assimilated [ʒ]-initial words can have some degree of exceptionality, since the lexically-indexed constraints associated with both subsets of [ʒ]-initial words are shown to have some weight depending on the context following [ʒ]. When disregarding the lexically-indexed constraints, a variable outcome is predicted for nonce words; that is, the context following [ʒ] determines the probability of assimilation. The tableau in (9) shows the predictions for nonce words in the vowel, schwa and consonant contexts.

(9)

		*l[cor]$_{\text{DEF}}$ 12.5	*l[cor] 0.02	Max(lat) 5.9	*#ʒʒ 5.9	Dep$_{\text{DEF}}$ 6.7	Dep 8.7	*CCC 24	Ident(cor) 15.4	*ə]$_{\sigma}$ 11.2	*#CCə 0.9	$\mathscr{H}$	p
/l$_{\text{DEF}}$/ + /ʒin/	ʒʒin	0	0	−1	−1	0	0	0	0	0	0	−11.9	≈.66
	lʒin	−1	−1	0	0	0	0	0	0	0	0	−12.6	≈.34
	ləʒin	0	0	0	0	−1	−1	0	0	−1	0	−26.7	≈0
/l$_{\text{DEF}}$/ + /ʒərq/	ʒʒərq	0	0	−1	−1	0	0	0	0	0	0	−11.9	≈.84
	lʒərq	−1	−1	0	0	0	0	0	0	0	−1	−13.5	≈.16
	ləʒərq	0	0	0	0	−1	−1	0	0	−1	0	−26.7	≈0
/l$_{\text{DEF}}$/ + /ʒrafa/	ʒʒrafa	0	0	−1	−1	0	0	0	0	0	0	−11.9	≈.97
	lʒrafa	−1	−1	0	0	0	0	−1	0	0	0	−36.6	≈0
	ləʒrafa	0	0	0	0	−1	−1	0	0	0	0	−15.5	≈.03

It can be seen from these tableaux that, when a full vowel follows [ʒ], assimilation is predicted 66% of the time. When a schwa follows [ʒ], assimilation is predicted 84% of the time. When a consonant follows [ʒ], assimilation is predicted 97% of the time. A comparison of the predictions of the lexicon and the MaxEnt learner incorporating lexically-indexed constraints about the behavior of [ʒ]-initial nonce words are shown in Table 4.

Context	**MaxEnt with LIC**	**Lexicon**
Vowel after [ʒ]	66%	37%
Schwa after [ʒ]	84%	81%
Consonant after [ʒ]	97%	96%

Table 4: Comparison of Predicted Probability of Assimilation in Each Context for Nonce Words between MaxEnt with lexically-indexed constraints and the Lexicon

3.7 *Discussion* As shown in Table 4, the overall assimilation patterns predicted by the MaxEnt learning model closely match the patterns observed in the lexicon, particularly for the schwa and consonant contexts. However, the predicted assimilation rate for the vowel context was significantly higher than the lexical statistics. Similar behavior has been observed by Hughto et al. (2019) in their investigation of exceptional

and variable patterns using a MaxEnt model with lexically-scaled constraints. Hughto et al. (2019) aimed to model both variation and exceptionality in four toy languages based on Russian vowel deletion. In Russian, the vowel of a CV prefix is deleted when attached to stems beginning with a vowel or a single consonant; when the stem begins with a CC sequence, vowel deletion is variable and lexically conditioned. One of the key findings in their study is the influence of majority patterns in the training data on the model's predictions. Hughto et al. (2019) show that, as the percentage of triggering CC-stems in the training data increases, the model predicts a higher probability of deleting the prefix vowel before any CC-stem, following the dominant pattern more strongly. This effect is especially prominent when the majority pattern is 60%-100% of the data. As a result, the model generalized this behavior to nonce forms. On the other hand, when there is no clear majority pattern, the model's predictions for nonce forms more closely followed the lexical statistics.

Similar to what has been shown by Hughto et al. (2019), one possible explanation for the higher assimilation rates predicted by the MaxEnt model for the vowel-following condition is the pressure from other conditions where assimilation occurs frequently or even categorically. First, there is an overall high proportion of assimilation among [ʒ]-initial words in the training data (62%). Second, all items beginning with non-[ʒ] coronals in the training data are predicted to assimilate categorically. This clearly shows that the majority patterns in the training data favors assimilation, explaining the model's higher probability of assimilation overall and, specifically, in the vowel context.

The next section presents the results of a nonce word experiment that tests the predictions of the MaxEnt learning model by investigating how Moroccan Arabic speakers generalize their knowledge about the assimilation patterns of [ʒ]-initial words to nonce forms.

4 Nonce Word Experiment

To test the predictions of the MaxEnt learning model, a forced-choice acceptability judgment experiment was conducted. The experiment was designed to explore the influence of the context following [ʒ] (whether a full vowel, a schwa, or a consonant follows [ʒ]) on the assimilation of nonce word. Participants were presented with assimilated and non-assimilated versions of each nonce word and were asked to choose their preferred version. The findings show that participants are sensitive to the phonological context following [ʒ], with a higher rate of assimilation when a consonant follows [ʒ], a lower assimilation rate when a vowel follows [ʒ], and an intermediate assimilation rate when a schwa follows [ʒ].

4.1 *Participants* In this experiment, 32 adult Moroccan Arabic speakers were recruited. Some of them were friends and family of the author, while others were recruited through word-of-mouth. The participants were at least 18 years old and were from the cities of Fes and Rabat in Morocco. The experiment was conducted entirely online where participants were able to complete the tasks at their convenience. On average, participants spent approximately 24 minutes to complete the experiment.

4.2 *Materials* The stimuli consisted of 42 words: 6 real words and 36 nonce words. Both types were chosen to represent the context following [ʒ]. Real words were equally divided into those with a consonant following [ʒ], those with a full vowel following [ʒ], and those with a schwa following [ʒ]. Among these, three words began with coronals, and three with non-coronals.

As shown in Table 5, nonce words were divided into three categories based on the sound following [ʒ] and were also divided into three sets depending on what kind of sound the word begins with. There were 14 words with a vowel following the initial consonant, among which 4 begin with non-coronals, 4 begin with non-ʒ coronals and 6 begin with [ʒ]. There were 14 words with a consonant following the initial consonant classified in the same manner to those with a vowel following the initial consonant. There were 8 words with a schwa following the initial consonant, among which 2 begin with non-coronals, 2 begin with non-ʒ coronals and 4 begins with [ʒ] coronals. No explicit hypotheses were formulated regarding the word shapes selected for each condition; the chosen word shapes were primarily selected due to their resemblance to existing Moroccan Arabic forms. For instance, the CCVC pattern is common for nouns in Moroccan Arabic, such as [ktab] 'book', [ħlib] 'milk', and [bnat] 'girls'. While the chosen nonce words were not close enough to any real words to be noticeably similar to participants, the consonant sequences selected are attested within Moroccan Arabic phonotactics.

Patterns		**non-coronals**	**non-ʒ coronals**	**[ʒ]-initial words**
CV	CVC	ħan, fux	saɣ, tuʒ	ʒuh, ʒin, ʒas
	CVCəC	fadər, ħarən	tikəl, nadəl	ʒuɣəm, ʒirəħ, ʒaləħ
CC	CCVC	xmig, kfax	zjal, ʃruf	ʒʕad, ʒmir, ʒbuq
	CCaCa	ħsama, xzada	ʃnara, zmada	ʒrafa, ʒmasa, ʒkala
Cə	CəCC	bərx	dənt	ʒərq, ʒəħt
	CəCCa	gəʃwa	nəhla	ʒəʕwa, ʒərqa

Table 5: The nonce words presented to participants

The words were presented in isolation without the need for a sentence frame, since assimilation is clearly observable in isolated forms. Both assimilated and non-assimilated versions of each item were recorded by the author using Praat (Boersma & Weenink, 2022) in a quiet environment. Before being presented to participants, the recorded items were examined by a native speaker of Moroccan Arabic who is unfamiliar with the task, and who judged the pronunciations to be natural and clearly assimilated or not.

4.3 *Procedure* The experiment was conducted using jsPsych (De Leeuw et al., 2023), a JavaScript framework for creating behavioral experiments. The script was developed and hosted on cognition.run, a platform designed by neuroscientists for running online experiments. The experiment consisted of two main parts, each designed to assess participants' preferences for assimilated versus non-assimilated versions of real and nonce words: the training trial where participants are trained on six real words (e.g. [ʕəwd] "horse") to get familiarized with the task, and the testing trial where participants were presented with the nonce words. In the training trial, participants were presented with audio recordings of six real words, each in both its assimilated and non-assimilated forms. These audio files were presented in a randomized order to prevent any potential order effects. The position of the assimilated and non-assimilated versions was also randomized. Participants listened to the two versions of each word and selected the one they preferred by clicking on a radio button adjacent to the audio clip. The question presented to participants was "which definite noun version you prefer?". The question was written in Moroccan Arabic (with Arabic script).

The testing trial of the experiment introduced the 36 nonce words, again in both assimilated and non-assimilated versions. Similar to the real words, participants listened to two versions of each nonce word and made their preference. Participants were forced to hear both audio clips and make a selection for one of the audio clips using radio buttons before the arrow button, which takes them to the following screen, became visible.

4.4 *Results* The results of the experiment followed the trends observed in the lexicon. In other words, the rates of assimilation for the definite article [l-] when followed by [ʒ]-initial nonce words were significantly influenced by the phonological context following [ʒ], i.e. whether a full vowel, a schwa, or a consonant followed [ʒ]. In the vowel-following context, the assimilation rate was observed at 66%. This rate increased to 74% in the schwa-following context, suggesting a stronger tendency to assimilate. The highest assimilation rate was 86% in the consonant-following context.

Statistical analysis was conducted using a logistic regression model to further assess the influence of the following context on the likelihood of assimilation as observed in the experimental results. The dependent variable was again binary (assimilation vs. no assimilation), with `FollowingSound` coded using Helmert contrasts to compare schwa vs. consonant and vowel vs. the average of schwa and consonant contexts. The analysis revealed significant effects of the context following [ʒ] on the probability of assimilation. In the schwa-following context, the likelihood of assimilation decreases ($\beta = -0.91$, $SE = 0.32$, $t = -2.79$, $p = 0.005$) compared to the consonant-following context. The likelihood of assimilation decreases further in the vowel-following context compared to both schwa and consonant contexts ($\beta = -0.97$, $SE = 0.24$, $t = -3.99$, $p < 0.0001$). Random intercepts for participants and items were included to account for variability across these factors.

4.5 *Discussion* The nonce word experiment revealed crucial insights about Moroccan Arabic speakers' generalization of the assimilation patterns. The results show that the likelihood of assimilation of [ʒ]-initial nonce words depends on the context following [ʒ]. Speakers showed variable behavior, which aligns with the predictions of the MaxEnt learner incorporating lexically-indexed constraints. A comparison of the predictions of the lexicon, the experiment, and the MaxEnt learner incorporating lexically-indexed constraints about the behavior of [ʒ]-initial nonce words are shown in Figure 1.

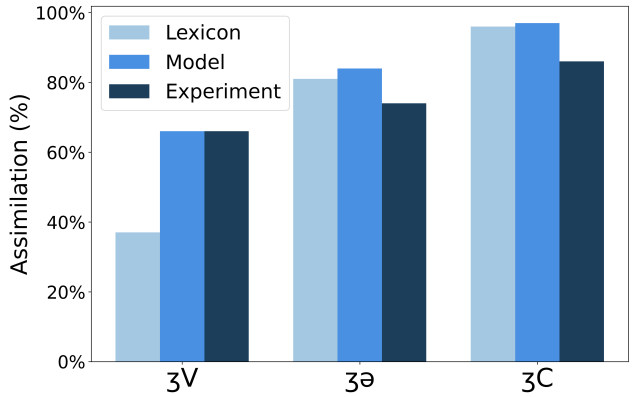

Figure 1: Comparison of predicted probability of assimilation in each context for nonce words between the lexicon, the MaxEnt model, and the experiment.

The assimilation rates for the consonant and schwa conditions in the experiment were, to some extent, close to the lexicon rates, i.e. a frequency matching behavior. However, the assimilation rate in the vowel context was significantly higher than the lexicon. This outcome may be due to certain aspect of the experimental design itself. Alternatively, it may be related to the organization of the lexicon/corpus. While the corpus used is representative of the Moroccan Arabic speaker's knowledge of [ʒ]-initial words, it is possible that Moroccan Arabic speakers use two distinct lexicons: one lexicon contains fully integrated (inherent) Moroccan words, while the other consists of words not fully assimilated into Moroccan Arabic. This latter set of words often resembles MSA words and is often associated with more educated, religious and political discourse.

To accurately determine if a word is inherently Moroccan Arabic, we must examine whether it has undergone vowel reduction/deletion. When words are derived from MSA, they undergo these phonological changes: short vowels are deleted (V → Ø), and long vowels are shortened (VV → V) (Kaye, 1987; Scheer, 1997). Words that follow this pattern, as shown (10), are considered inherent Moroccan Arabic words. If we only consider such words in the corpus, the assimilation rate in the vowel context increases to 65%, closely aligning with the experimental results. Therefore, it is possible that, when predicting the behavior of nonce [ʒ]-initial words with a following vowel, participants used the lexical statistics associated with a sub-lexicon composed exclusively of Moroccan Arabic inherent words.

(10)

MSA	**Moroccan Arabic**	**Gloss**
kalaam	klam	'speech'
ʒaar	ʒar	'neighbor'

5 Conclusion

This study investigated the assimilation patterns of the Moroccan Arabic definite article [l-] focusing on the variation and exceptionality observed in [ʒ]-initial words. By examining a comprehensive corpus and nonce word experimental data, this study challenged the previously proposed binary categorizations about the assimilation of [ʒ]-initial words and showed that the observed productive and exceptional patterns are gradient. The findings reveal that assimilation of [ʒ]-initial words is not a categorical but rather a variable phenomenon influenced by the phonological context following [ʒ]. It has also been shown that a MaxEnt

model with lexically-indexed constraints is successful in learning the exceptionality and variation observed within the definite article assimilation patterns. This model was able to account for the observed variability in [ʒ]-initial nonce words as well as the fixed pronunciations of existing [ʒ]-initial words. In terms of exceptionality, the model treated both assimilated and non-assimilated [ʒ]-initial real words as exceptional.

The findings from this study contribute to the ongoing discussions regarding the nature of phonological representations and the mechanisms through which phonological patterns are learned. The definite article assimilation patterns in Moroccan Arabic support the view that phonological processes can be gradient and influenced by lexical statistics. This provides an argument for the probabilistic nature of phonological knowledge and its representation. Unlike most previous studies that examine artificial language data or toy languages, this paper offered a detailed examination and application of the MaxEnt model incorporating lexically-indexed constraints to a realistic dataset.

One possible avenue for future research is to examine the predictions of alternative models for learning variable and exceptional patterns (Becker & Gouskova, 2016; Smolensky & Goldrick, 2016; Shih, 2018; Hughto et al., 2019) about the assimilation patters investigated in this paper as well as similar complex patterns. Another consideration that is worth revisiting is the predictions of the MaxEnt model with lexically-indexed constraints for the behavior of nonce words. Despite the commonly held assumption that such models predict frequency matching, we have seen that the model predicted an higher rate of assimilation for the vowel condition, compared to the lexical trends. This behavior has been acknowledged by previous studies (Moore-Cantwell & Pater, 2016; Hughto et al., 2019). However, further research is needed to reveal the factors influencing this behavior.

References

Al Ghadi, Abdellatif (1990). *Moroccan Arabic Plurals and the Organization of the Lexicon*. Faculty of Letters and Humanities, Mohammed V University.

Bates, Douglas, Martin Mächler, Benjamin M. Bolker & Steven C. Walker (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67:1, 1–48.

Becker, Michael & Maria Gouskova (2016). Source-oriented generalizations as grammar inference in russian vowel deletion. *Linguistic inquiry* 47:3, 391–425.

Benhallam, Abderrafi (1980). *Syllable Structure and Rule Types in Arabic*. Ph.D. thesis, University of Florida.

Boersma, Paul & Joe Pater (2016). Convergence properties of a gradual learning algorithm for harmonic grammar. *Harmonic grammar and harmonic serialism* 389–434.

Boersma, Paul & David Weenink (2022). Praat: doing phonetics by computer. `http://www.praat.org`. Version 6.2.09.

Boudlal, Abdelaziz (2001). *The Prosody and Morphology of a Moroccan Arabic Dialect: An Optimality-Theoretic Account*. VDM Verlag, Rabat. Issue: March.

De Leeuw, Joshua R., Rebecca A. Gilbert & Björn Luchterhandt (2023). jsPsych: Enabling an open-source collaborative ecosystem of behavioral experiments. *Journal of Open Source Software* 8:85, p. 5351, URL `https://joss.theoj.org/papers/10.21105/joss.05351`.

Ernestus, Miriam & R. Harald Baayen (2003). Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language* 79:1, 5–38.

Freeman, Aaron (2016). Arabic j and the class of Sun Letters: A historical and dialectological perspective. *Perspectives on Arabic Linguistics XXVII*, John Benjamins, 171–185.

Goldwater, Sharon & Mark Johnson (2003). Learning OT constraint rankings using a maximum entropy model. *Proceedings of the Workshop on Variation Within Optimality Theory*, Stockholm University, 111–120.

Harrell, Richard (1962). *A Short Reference Grammar of Moroccan Arabic*. Georgetown University Press.

Harrell, Richard & Harvey Sobelman (1966). *A Dictionary of Moroccan Arabic: Arabic-English*. Georgetown University Press, Washington, D.C.

Hayes, Bruce (2009). *Introductory Phonology*. Wiley-Blackwell.

Hayes, Bruce & Zsuzsa Cziráky Londe (2006). Stochastic Phonological Knowledge: The Case of Hungarian Vowel Harmony. *Phonology* 23:1, 59–104. Publisher: Cambridge University Press.

Hayes, Bruce & Colin Wilson (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic inquiry* 39:3, 379–440.

Heath, Jeffrey (1987). *Ablaut and Ambiguity: Phonology of a Moroccan Arabic dialect*. State University of New York Press, Albany.

Heath, Jeffrey (1989). *From code-switching to borrowing: foreign and diglossic mixing in Moroccan Arabic*. Library of Arabic linguistics ; monograph no. 9, Kegan Paul International, London.

Hughto, Coral, Andrew Lamont, Brandon Prickett & Gaja Jarosz (2019). Learning exceptionality and variation with lexically scaled maxent. *Society for Computation in Linguistics* 2:1.

Kaye, Jonathan (1987). Government in Phonology. The Case of Moroccan Arabic. *The Linguistic Review* 6:2.

Legendre, Géraldine & Paul Smolensky (2006). *The harmonic mind : from neural computation to optimality-theoretic grammar*. MIT Press.

Legendre, Géraldine, Yoshiro Miyata & Paul Smolensky (1990a). Harmonic grammar – a formal multi-level connectionist theory of linguistic well-formedness: An application. *12th Annual Conference of the Cognitive Science Society*, Psychology Press.

Legendre, Géraldine, Yoshiro Miyata & Paul Smolensky (1990b). Harmonic grammar – a formal multi-level connectionist theory of linguistic well-formedness: Theoretical foundations. *Proceedings of the 12th Meeting of the Cognitive Science Society*.

Linzen, Tal, Sofya Kasyanenko & Maria Gouskova (2013). Lexical and phonological variation in russian prepositions. *Phonology* 30:3, 453–515.

Maas, Utz & Stephan Procházka (2022). Nominal determination in Moroccan Arabic:. *Studies in Language* 46:4, 793–846. Publisher: John Benjamins Publishing Company.

Moore-Cantwell, Claire & Joe Pater (2016). Gradient exceptionality in maximum entropy grammar with lexically specific constraints. *Catalan Journal of Linguistics* 15, 53–66.

Outchakoucht, Aissam & Hamza Es-Samaali (2021). Moroccan dialect -darija- open dataset.

Pater, Joe (2000). Non-uniformity in English secondary stress: the role of ranked and lexically specific constraints. *Phonology* 17:2, 237–274. Publisher: Cambridge University Press.

Pater, Joe (2009). Morpheme-Specific Phonology: Constraint Indexation and Inconsistency Resolution Publisher: Equinox Publishing Ltd.

Pater, Joe, Robert Staubs, Karen Jesney & Brian Cantwell Smith (2012). Learning probabilities over underlying representations. *Proceedings of the twelfth meeting of the Special Interest Group on Computational Morphology and Phonology*, 62–71.

R Development Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, URL `http://www.r-project.org`.

Ridouane, Rachid (2016). Leading issues in tashlhiyt phonology. *Language and Linguistics Compass* 10:11, 644–660.

Scheer, Tobias (1997). Vowel-zero alternations and their support for a theory of consonantal interaction p. 67. Publisher: Rosenberg & Sellier.

Shih, Shu-hao (2018). On the existence of sonority-driven stress in Gujarati. *Phonology* 35, 327–364.

Smolensky, Paul & Matthew Goldrick (2016). Gradient symbolic representations in grammar: The case of french liaison. *Rutgers Optimality Archive* 1552, 1–37.

Staubs, Robert (2011). Harmonic grammar in r. Software package for studying Harmonic Grammar and Maximum Entropy Grammar in R, including features for hidden structure learning. Available at: `https://websites.umass.edu/hgr/`.

Zuraw, Kie (2000). *Patterned exceptions in phonology*. Ph.D. thesis, UCLA.