**Investigation of the Distributions, Derivation, and Generalizations in Arabic Plural System**

by

Fahad Hamad A. Alrashed

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Linguistics)
in the University of Michigan
2021

Doctoral Committee:

Professor Jeffrey G. Heath, Chair
Associate Professor Steven P. Abney
Professor Mohammad Alhawary
Professor San Duanmu

Fahad Hamad A. Alrashed

fahadal@umich.edu

ORCID iD:  0000-0003-0977-7011

**Acknowledgements**

This project would not have been possible without the help and support of many people. I thank them all, but I want to single out some of them here.

First and foremost, I want to express my profound gratitude to my academic advisor, Dr. Jeff Heath, who has been involved in this work from the very beginning and has, for years to come, patiently guided my steps through the winding roads of Arabic morphophonology. Dr Heath offered support and encouragement without which I doubt the work would have ever seen the light.

I also want to thank all other members of my committee who have been instrumental in the development of this dissertation: Dr. Steve Abney, Dr. Mohammad Alhawary, and Dr. San Duanmu. I thank them for their time, their constructive criticism, and guidance.

I would also like to extend my thanks to my family: my parents, my brothers, my sisters for supporting and encouraging me throughout my studies.

Last, but certainly not least, my continuous thanks go to my wife, Maha, for her unflinching support and everlasting love. This dissertation would not have been possible without her.

## Table of Contents

## List of Tables

## List of Figures

**Abstract**

The Arabic plural system poses a challenge to current morphological accounts since the regular sound plural that is formed by suffixation contrasts with irregular broken plurals formed by internally modifying the singular stem. Although aspects of the Arabic plural system have been widely studied since the early ages of Arab grammarians (Abu Al-Saud 1971; Yaaqub 2004), there are several issues that remain unresolved and warrant further investigation. This dissertation uses a combination of statistical, qualitative, and computational approaches to provide a comprehensive account of several outstanding problems in Arabic nominal plurals.

Theoretical investigations of Arabic nominal plurals have led to conflicting results about the status of Arabic nominal ablaut as a minority default system (McCarthy & Prince 1990; Boudelaa & Gaskell 2002). Apart from this, little has been said about other aspects of the distribution of Arabic plurals, namely the interplay between regularity of the plural and its frequency in actual language use (Bybee 2001). This dissertation takes a usage-based approach to revisit the question of the status of Arabic as a minority-default system and to examine the interplay between the productivity and the frequency in actual language use of plural types. The results from the statistical distribution of sound and broken plurals are in line with the claim Arabic pluralization is a minority default system. The results are also consistent with the prediction made by the usage-based model that low type frequency compensates for weak lexical strength by high token frequency.

The dissertation also investigates the role of singular stem weight on plural derivation. Numerous attempts have been made to model Arabic broken plurals, which fall into three main

groups according to their specific morphological approach: Generative Morphophonology (Brame 1970; Levy 1971), Root-&-Pattern Morphology (McCarthy 1979; Hammond 1988), and Prosodic Morphology (McCarthy and Prince 1990). However, there has not been any investigation of the influence of the additive weight of singular stems on the derivation of plural forms. Results from qualitative and computational analyses provide evidence for the role of simple additive weight on the mapping of singular input stems to plural outputs in Arabic broken plural. Stem weight does not completely determine the plural pattern. Rather, its role can be viewed as a (quasi) well-formedness condition on plural templates based on input forms.

The dissertation examines the types of information that are relevant to the Arabic plural system by performing a computational analysis on singular-plural pairs collected from a comprehensive corpus. The performance of multiple K-Nearest Neighbor (KNN) classifiers that use different combinations of factors to select plural patterns are compared to determine the importance of each factor. The results show that the CV template, vowel melody and semantic qualities of the singular all contribute to determining the shape of the plural template, though with varying degrees. The syllabic shape of the singular forms of Arabic nouns is the major factor in predicting their plural forms, followed by the vowel melody and the semantic features.

# Chapter 1 Introduction

## 1.1 Objectives

Perhaps the most widely acknowledged contribution that the study of Arabic has already made to the theory of linguistics has been in the areas of non-linear morphology (Comrie 1991; Versteegh 1997). Arabic, and more generally the Semitic languages, were well known, even before the current interest in the study of synchronic language, for expressing morphological derivation not through affixes but through internal modification of the word. One of the issues that has gained an enormous interest among researchers in Arabic nonlinear morphology is the problem of the Arabic plural system. The structure of the plural in Arabic is characterized by complexity and increased allomorphy, which has been instrumental in the development of recent nonlinear approaches to morphology such as the autosegmental (Root-and-Pattern) theory (McCarthy 1979; Hammond 1988) or the theory of prosodic morphology (McCarthy and Prince 1990). Although aspects of Arabic plural system have been widely studied since the early ages of Arab Grammarians' tradition (Abu Al-Saud 1971; Yaaqub 2004), there are several issues that remain unresolved and warrant further investigation. In this dissertation, I intend to address three main issues: the distribution of plural patterns, the influence of stem weight on the mapping of singular nouns to plural forms, and the factors that determine the mapping between singular stems and plural forms.

Although many studies have analyzed the statistical distribution of the sound and broken plural in the Arabic plural system, these studies have serious limitations regarding the data used

in the analysis. The majority of these studies are based on data collected from one dictionary of literary Arabic. The few studies that did not rely on dictionary source for the data have used a relatively small corpus. Consequently, these studies drew conclusions about the Arabic plural system based on data that is not representative of actual use of the language. The distribution of nominals between the two types of plural remains one of the problems in Arabic. Without using a large corpus that is representative of actual use of Arabic, inquiries about the dominant type of plural in a representative corpus and the interaction between regularity and productivity remain unknown. One of the goals of this dissertation is to address this gap in the literature.

The role of stem weight in the mapping of the singular stems to their plural forms is another missing piece of research in the area of non-linear approaches to Arabic morphology. In the current work on problems in Arabic morphology, some progress has been made toward developing models which would highlight the formal relationship between derived forms and their sources. These models fall into three groups based on their theoretical framework: rule-based models that employ derivational rules in an explicitly generative framework, a templatic model that combines the root + pattern analysis with principles of autosegmental phonology, and a prosodic model that incorporates principles of prosodic analysis to morphology. However, none of these models, including the one that uses a prosodic approach, refer to the role of the stem weight or size in the mapping of the singular stems to their plural forms. A goal of this dissertation is to develop a model of Arabic plural system that incorporates the role of the weight stem in process of mapping the singular stems to their plural forms.

The factors that determine morphological learnability and productivity in Arabic plural system are another area of research that needs further investigations. Current accounts of the aspects of morphological productivity in the nominal plural system in Arabic are challenged by

the system complexity represented by the number of plural patterns and the less understood mapping rules of singular to plural. Understanding aspects of productivity and regularity in the Arabic nominal plural is important to determine how morphological learnability and generalizations are governed in the system. However, these aspects cannot be fully understood without identifying the importance of factors that determine morphological regularity and productivity in the system. While there are a few studies that have examined the learnability of nominal plurals in Arabic using computational models, they give little insight into the sources of information that the models utilize to predict a plural pattern for a given singular stem. This dissertation aims at developing a computational predictive model to provide a comprehensive account of the factors that govern the learnability and productivity in the Arabic plural system.

## 1.2 Modern Standard Arabic

The analysis presented in this dissertation relies exclusively on data from Modern Standard Arabic (henceforth MSA). So, it is helpful to discuss this variety of Arabic and review its phonemic inventory. In this section I present some basic information about the phonological and morphological structure of Arabic as represented in most grammar textbooks and descriptive studies. I start by giving a description of the consonant system and the vowel system. The section ends with a brief description of the morphological system.

Arabic is the main language in the Arab countries which occupy most of the Middle East and North Africa. Close to 430 million people in that region speak one variety of Arabic or another as their first language. Ferguson (1959) considers the language situation in the communities where Arabic is the main language as representing a form of diglossia. Ferguson states that the characteristic feature of a diglossic language community is that the community has a variety that is exclusively for formal uses and is not derived, or based on, the natural spoken

variety. According to Ferguson, MSA in Arabic speaking community represents the 'high' or the standard variety while the natural spoken language or the regional vernacular Arabic is the related 'low' spoken variety. Many Arab intellectuals hail MSA as a more 'proper' form of Arabic than the regional vernaculars which they view as signs of the corruption that befell the revered Classical Arabic. MSA is currently the language of the media, the public education systems, practically all written and technical forms of Arabic, and intellectual circles. MSA can also be thought of as a pan-Arab lingua franca used whenever dialectal differences veer into unintelligibility. According to Holes (1994), the wide spreading of education and mass-media exposure has a "leveling influence" which brings the divergent Arabic dialects gradually closer to MSA.

Modern Standard Arabic has 28 consonants (given in Table 1-1). In cells with two consonants, the one on the left is voiceless while the one on the right is voiced.

| | Labial | Labio-dental | Inter-dental | Alveolar | Alveo-palatal | velar | uvular | pharyngeal | laryngeal |
|---|---|---|---|---|---|---|---|---|---|
| **Stop** | b | | | t, d | ʄ | k | q | | ʔ |
| **Nasal** | m | | | n | | | | | |
| **Trill** | | | | r | | | | | |
| **Fricative** | | f | θ, ð | s, z | ʃ | x, ɣ | | ħ, ʕ | h |
| **Approximant** | | | | | j | w | | | |
| **Lateral** | | | | l | | | | | |
| **Pharyngealized stop** | | | | tˤ, dˤ | | | | | |
| **Pharyngealized fricative** | | | ðˤ | sˤ | | | | | |

*Table 1-1.* MSA consonantal inventory.

Arabic consonants present phonemic contrasts in both articulatory manners and places of articulation. Arabic has a large number of places of articulation: five for stops (including glottal stops) against an average of three in other languages in the University of California-Los Angeles Phonological Segment Inventory Database (henceforth UPSID) (Maddieson 1984), and six places of articulation for fricatives against an average of four in other languages. It is true that Arabic has a rich consonantal inventory. It is also the case that this inventory is characterized by

several gaps. The phonemic gaps are illustrated by the lack of voicing contrasts for the following consonants: /b f k q ðˤ sˤ ɟ ʃ/.

Modern standard Arabic has a limited vocalic system that consists of three cardinal vowels /i u a/. Vowel duration is contrastive in Arabic so each of the three vowel quality can appear in short vs. long forms in minimal pairs. There are two diphthongs made by combining the low vowel with glides: /aj/ and /aw/. The vocalic inventory is given in Table (1-2).

| | Front | Central | Back round |
|---|---|---|---|
| High | i, ii | | u, uu |
| Low | | a, aa | |

*Table 1-2*. MSA vowel inventory.

Many textbooks and descriptive grammars (e.g. Watson 2002, Holes 2004) describe the nonlinear formation of words in MSA as displaying a templatic morphology or Root-and-Pattern morphology. In this analysis, all words consist of two elements: a consonantal root and a pattern or template. The first part, the consonantal root, consists of a sequence of consonants that gives a general meaning or concept. Words that share the same consonantal roots have meanings that are broadly related to the same concept. For example, the root k-t-b denotes a general concept of writing, and words that are generated from this root revolve around the concept of writing. The most common consonantal roots consist of three consonants. There are also biliteral roots that have two consonants, and quadriliteral roots consisting of four consonants.

The second element in Arabic word formation is the pattern which consists of a syllabic template (or CV template) with one or more vowels and, in some cases, one or more consonants. While the abstract consonants in the root are said to convey a general meaning or concept, the patterns are argued to convey grammatical meanings. For example, Holes (2004) said that the combination of the template $C_1vC_2vC_3$ with the vowel set of a-a denotes an action, transitive or

intransitive, performed by an agent. So, by inserting the consonantal roots k-t-b into the pattern *CaCaC*, we get the verb /katab/ "he wrote". To sum up, the lexical meaning of the root combines with the grammatical meaning expressed by the pattern to generate a stem or base with the specific meaning. The following examples demonstrate how different words that belong to the concept of writing are created by inserting the root k-t-b into the different patterns:

| | **Form** | **Pattern** | **Gloss** |
|---|---|---|---|
| a. | katab | CaCaC | "he wrote" |
| b. | kattab | CaCCaC | "he caused (someone) to write" |
| c. | jaktub | jaCCuC | "he writes" |
| d. | jukattib | juCaCCiC | "he causes (someone) to write" |
| e. | kaatib | CaaCiC | "writer" |
| f. | kitaab | CiCaaC | "book" |
| g. | maktab | maCCaC | "office, desk" |
| h. | maktabat | maCCaCat | "library" |

*Table 1-3.* The formation of the words by inserting the radical consonants k-t-b in the specified patterns.

## 1.3 Dissertation outline

The dissertation is divided into 6 chapters as follows. In Chapter One I describe the objectives of the research and briefly describe the phonological and morphological aspects of the language of interest, specifically MSA. Chapter Two gives a basic description of the plural formation in Arabic. In Chapter Three I present and discuss the results of the analysis of the distribution of Arabic sound and broken plurals. In Chapter Four I discuss the role of additive weight of the singular stem on the observed singular-to-plural mapping in Arabic broken plural system. In Chapter Five I present the results of the computational analysis on the contribution of morphophonological and semantic factors in the mapping of singulars to plurals in Arabic plural system. Finally, concluding remarks and suggestions for future research are offered in Chapter Six.

# Chapter 2 The Arabic Plural System

The study of the Arabic plural system goes back to the days of traditional Arab Grammarians. According to recent reviews of this grammatical traditions by Abu Al-Saud (1971) and Yaaqub (2004), Classical Arabic recognizes three number categories: singular, dual and plural. They further divide the plural into 'sound plural' جمع سالم and 'broken plural' جمع تكسير. This distinction between the two types of plural was retained by many contemporary varieties of Arabic such as Egyptian, Yemeni (Watson 2004), Maghribi (Heath 1987) and MSA. This section will present a brief background of the morphological structure of the two plural types as they occur in MSA.

## 2.1 The Sound Plural

As the name suggests, the singular stem in the sound plural remains intact, and the plural is formed simply by concatenating suffixes to the singular stem. In addition to number, Arabic nouns are marked for case and gender. There are three cases, namely accusative, genitive and nominative, and two genders, masculine and feminine. Traditional Arab Grammarians list the following set of suffixes to represent this distinction:

| | **Case** | **M.Sg.** | **M.Pl.** | **F.Sg.** | **F.Pl.** |
|---|---|---|---|---|---|
| a. | Nom. | *-un* | *-uuna* | *-atun* | *-aatun* |
| b. | Gen. | *-in* | *-iina* | *-atin* | *-aatin* |
| c. | Acc. | *-an* | *-iina* | *-atan* | *-aatin* |

*Table 2-1.* Sound plural suffixes. There are three case categories: Nominative (Nom.), Genitive (Gen.) and Accusative (Acc.), two number categories: Singular (Sg.) and Plural (Pl.), and two genders: Masculine (M.) and Feminine (F.).

Table 2-1 lists ten stem-external suffixes which indicate case, gender and number categories for a given noun. A plural form is created by augmenting the singular stem with the plural suffix that corresponds to the case and gender of that stem. For example, the masculine nominative singular noun [muhandis-un] “engineer” becomes plural by adding the masculine nominative plural suffix [-uuna] to that singular noun, yielding [muhandis-uuna].

As indicated in Table 2-1, the traditional analysis of Arabic nouns treats the process of sound pluralization as addition of stem-external suffixes, where each suffix corresponds to a combination of case, number and gender categories. Ratcliff (1990, 1998) argues that it is possible to decompose these ten suffixes into three distinct morphemes each of which correlates one-to-one with a specific functional category. According to Ratcliff, masculine gender is the default and feminine gender is marked by a segmental affix [-t]; case is marked by mapping a vowel quality to the V segment most peripheral to the stem; and plural number is marked by affixing a timing unit (V slot) to the right stem boundary.

## 2.2 The Broken Plural

The broken plural, on the other hand, is formed by breaking the stem. The shape of the stem is altered by a non-concatenative process. Vowels, and sometimes special consonants, are inserted between the preserved consonants of the singular stem in accordance with a specific plural template. There are more than 30 broken plural patterns in Arabic. Wright (1971) lists 31 broken plural patterns. McCarthy and Prince (1990) in their analysis of the Arabic broken plural divided Wright’s (1971) 31 templates into 4 prosodically modified groups, Table (2-2):

| **a. Iambic** | **b. Trochaic** | **c. Monosyllabic** | **d. Other** |
|---|---|---|---|
| *CiCaaC* | *CuCaC* | *CuCC* | *CuC$_j$C$_j$aC* |
| *CuCuuC* | *CiCaC* | *CiCC + at* | *CuC$_j$C$_j$aaC* |
| *CaCaaC* | *CaCaC* | *CiCC + aan* | |
| */CaCaaC/* surfacing as *ʔaCCaaC* | *CiCaC + at* | *CuCC + aan* | |
| *CaCaaC + /ay/* | */CaCuC/* surfacing as *ʔaCCuC* | *CaCC + /ay/* | |
| *CaCiiC* | *CuCuC* | *CaCC* | |
| *CuCuuC + at* | *CaCaC + at* | | |
| *CiCaaC + at* | *CuCaC + at* | | |
| *CawaaCiC* | *CuCaC + aaʔ* | | |
| *CaCaaʔiC* | */CaCiC/ + at* surfacing as *ʔaCCiCat* | | |
| *CaCaaCiC* | */CaCiC/ + aaʔ* surfacing as *ʔaCCiCaaʔ* | | |
| *CaCaaCiiC* | | | |

*Table 2-2.* The four groups of the most frequent plural patterns from McCarthy and Prince based on Wright.

The 31 plural templates in Table (2-2) happen to be the most frequent patterns of broken plurals. Plunkett and Nakisa (1997), however, argue that when the infrequent patterns are included, the number is probably greater than 70. When the infrequent templates, such as those that occurred less than ten times in their data-set, are removed, the number of plural templates is reduced from 71 to 12 patterns. The most frequent plural forms in their data-set are in Table (2-3):

| **Pattern** | **Frequency** | **% of total** |
|---|---|---|
| *CaCaaCiC* | 150 | 17.46 |
| *aCCaaC* | 140 | 16.30 |
| *CuCuuC* | 83 | 9.66 |
| *aCaaCiiC* | 58 | 6.75 |
| *CiCaaC* | 42 | 4.89 |
| *CuCaC* | 32 | 3.79 |
| *CuCaCaaC (adjectival)** | 29 | 3.38 |
| *CaCaCaaC* | 29 | 3.38 |
| *CuCaaC ** | 24 | 2.79 |
| *CuCuC* | 21 | 2.44 |
| *aCCiCa* | 20 | 2.33 |
| *CiCaC* | 14 | 1.63 |

*Table 2-3.* Frequent Plural patterns from Plunkett and Nakisa (1997).

In addition to distinguishing the broken plural from the linear sound plural, traditional Arab grammarians further divide the broken plural patterns in McCarthy and Prince (1990) and

Plunkett and Nakisa (1997) into several categories based on the quantity that the plural noun denotes. The first category is the plural of paucity, which is used to denotes a group of entities ranging between three to ten. According to Abu Al-Saud (1971) and Yaaqub (2004), Arab grammarians specify four patterns that denote plural of paucity. These patterns are illustrated in Table (2-4):

| | **Pattern** | **Plural** | **Singular** | **Gloss** |
|---|---|---|---|---|
| a. | *ʔaC₁C₂uC₃* | *ʔabħur* | *baħr* | *"sea"* |
| b. | *ʔaC₁C₂iC₃at* | *ʔasliħat* | *silaaħ* | *"weapon"* |
| c. | *ʔaC₁C₂aaC₃* | *ʔaɟsaam* | *ɟism* | *"body"* |
| d. | *C₁iC₂C₃at* | *ɣilmat* | *ɣulaam* | *"boy"* |

*Table 2-4.* Patterns that designate plural of paucity

The second category of broken plurals is the plural of multiplicity. Like the plural of paucity, this category is used when referring to a specific quantity, namely a group that exceeds ten. Yaaqub (2004) reports that Arab grammarians do seem to have a consensus on the number of patterns in this category. Basically, they consider any pattern that does not fit in the other broken plural categories to be part of the plural of multiplicity. Abu Al-Saud (1971) lists up to 30 patterns that belong to the plural of multiplicity. Table 2-5 shows some of these patterns:

| | **Pattern** | **Plural** | **Singular** | **Gloss** |
|---|---|---|---|---|
| a. | *C₁uC₂uC₃* | *rusul* | *rasuul* | *"messenger"* |
| b. | *C₁uC₂aC₃* | *ɣuraf* | *ɣurfat* | *"room"* |
| c. | *C₁aC₂aC₃at* | *katabat* | *kaatib* | *"writer"* |
| d. | *C₁ uC₂C₂aC3* | *suɟɟad* | *saaɟid* | *"boy"* |
| e. | *C₁iC₂aaC₃* | *θiyaab* | *θawb* | *"dress"* |
| f. | *C₁uC₂uuC₃* | *quluub* | *qalb* | *"heart"* |
| g. | *C₁uC₂aC₃aaʔ* | *kuramaaʔ* | *kariim* | *"generous"* |

*Table 2-5.* Patterns denoting plural of multiplicity.

The last category of the broken plurals is the ultimate plural. This is for plurals that do not allow further pluralization. For example, Yaaqub (2004) lists the example [kalb] 'dog' which has two different plural forms as in [kilaab] and [ʔaklub], and contrasts this example with

[ɟawaamiis], the plural form for a word [ɟaamuus] ‘buffalo’, which does not allow any further pluralization. Thus, Arab grammarians refer to this type of broken plural as the ultimate plural. The ultimate plural patterns are listed in Table 2-6:

| | Pattern | Plural | Singular | Gloss |
|---|---|---|---|---|
| a. | *$C_1$awaa$C_2$i$C_3$* | *xawaatim* | *xaatim* | *“ring”* |
| b. | *$C_1$awaa$C_2$ii$C_3$* | *ɟawaamiis* | *ɟaamuus* | *“buffalo”* |
| c. | *$C_1$a$C_2$aaʔi$C_3$* | *xamaaʔir* | *xamiirat* | *“yeast”* |
| d. | *$C_1$a$C_2$aa$C_3$iij* | *sˁaħaariij* | *sˁaħraaʔ* | *“dessert”* |
| e. | *$C_1$a$C_2$aa$C_3$i$C_4$* | *ʕaqaarib* | *ʕaqrab* | *“scorpion”* |
| f. | *$C_1$a$C_2$aa$C_3$ii$C_4$* | *salaatˁiin* | *sultˁaan* | *“sultan”* |

*Table 2-6.* Patterns linked with the ultimate plural.

In spite of the fact that the number of broken plural patterns in Arabic is finite, it is not always easy to predict which of the attested broken plural patterns a singular stem will select. Stems that have the same shape in the singular do not necessarily take the same plural pattern. For example, all the words in (2-7) below have the same *CvCC* shape in the singular, but they all take different broken plural patterns:

| | Singular | Plural | Plural Pattern | Gloss |
|---|---|---|---|---|
| a. | *qalb* | *quluub* | *CuCuuC* | *“heart”* |
| b. | *kalb* | *kilaab* | *CiCaaC* | *“dog”* |
| c. | *lawn* | *ʔalwaan* | *ʔaCCaaC* | *“color”* |
| d. | *qird* | *qiradat* | *CiCaCat* | *“monkey”* |
| e. | *θawr* | *θiiraan* | *CiiCaaC* | *“bull”* |

*Table 2-7.* Singular stems with the same syllabic shape map to different plural patterns.

Conversely, words that have the same plural pattern do not necessarily have the same shape in the singular. For example, all the words in (2-8) below take the plural pattern *ʔaCCaaC*, but each has a different shape in the singular:

| | Plural | Singular | Singular Pattern | Gloss |
|---|---|---|---|---|
| a. | *ʔaqlaam* | *qalam* | *CvCvC* | *"pen"* |
| b. | *ʔaħfaad* | *ħafiid* | *CvCvvC* | *"grandchild"* |
| c. | *ʔabqaar* | *baqarat* | *CvCvCvt* | *"cow"* |
| d. | *ʔasˁħaab* | *sˁaaħib* | *CvvCvC* | *"friend"* |
| e. | *ʔaθwaab* | *θawb* | *CvCC* | *"dress"* |
| f. | *ʔanyaab* | *naab* | *CvvC* | *"tusk"* |

*Table 2-8.* Plural patterns with the same shape are linked with singular stems with different shapes.

The examples in Table (2-7) and (2-8) show that when the syllabic shapes associated with the singular and the plural are considered, the morphological structure of Arabic broken system is characterized by a large amount of many-to-many mappings between singulars and plurals and overlaps within singular or plural classes. In fact, the redundant alternation as a result of the many-to-many mappings is one of the main problems of broken plurals in Arabic. Arab grammarians have dealt with the problem of redundant forms and alternations in Arabic broken plurals. Their approach to the alternation problem is to use the semantic functions associated with the plural patterns (plural of paucity, plural of multiplicity, etc.) to account for the redundant plural forms. So, when a singular noun has two plural forms, Arab grammarians will use the semantic distinction to distinguish between the redundant forms. One will be for plural paucity while the other will be for the plural multiplicity. Levy (1971) argues that there is no statistical or systematic data that support this argument that the redundant plural forms for a particular singular serve a semantic function.

# Chapter 3 Analysis of the Distribution of Arabic Sound and Broken Plurals

## 3.1 Introduction

In this part of the dissertation, I want to focus on the frequency distribution of the sound and broken plural in Arabic plural system. The distribution of nominals between the two types of plural remains one of the problems in Arabic. Inquiries about the dominant type of plural in a representative corpus and the interaction between regularity and productivity remain unknown. Analyzing the statistical frequency of the plural types can help address these problems. In the current study, we address these issues by subjecting data collected from a corpus of over 300,000,000 words of Arabic text compiled from written sources in Gigaword, a comprehensive corpus of Modern Standard Arabic, to answer these questions.

## 3.2 Broken plural patterns and the Default system of pluralization in Arabic

As explained in Chapter 2, the sound plural in MSA that is formed by suffixation of singular stem competes with approximately 31 broken plural patterns that a singular stem may select. Because they involve a non-concatenative morphological process, broken plurals are considered in Arabic literature as the irregular form of pluralization. However, it is argued that broken pluralization is very pervasive throughout the system. According to early several surveys and studies by Levy, (1971), Murtonen, (1964), Wright (1971) and McCarthy and Prince (1990) that investigated the statistical distribution of nominals by plural type, broken plural constitutes the main process of plural formation for the majority of nouns. Another area where the pervasiveness of the broken plural is seen is loan-words, many of which form plurals by taking

broken plural patterns. For example, the plural for the loan-words [*bank*] (bank) and [*kart*] (card) are [bunuuk] and [kuruut]. The plural is formed by changing the vowel [*a*] in the singular stem to [*u*], and then inserting [*uu*] between the second and the third consonant.

The idea that the broken plural is the default inflectional marker of pluralization and the sound plural is just applied to a minority group of nouns became a common theme among the studies that followed the early surveys. Most of the credit (or blame depending on where one stands on the issue) of spreading the idea of Arabic having a minority default system of pluralization is attributed to McCarthy and Prince's (1990) work on the problem of Arabic broken plural. According to them "essentially all canonically-shaped lexical nouns of Arabic take broken plurals", while the sound plural is "systematically found only with the following short list: proper names; transparently derived nouns or adjectives such as participles, deverbals and diminutives; non-canonical or unassimilated loans and the names of the letters of the alphabet" (McCarthy & Prince, 1990: p. 212). Plunkett and Nakisa (1997) put this claim by McCarthy and Prince to test by counting the percentage of sound and broken plurals occurring in a selection of approximately 1000 nouns randomly taken from the Wehr (1976) dictionary, the same dictionary McCarthy & Prince used to do their analysis. Confirming McCarthy and Prince, Plunkett and Nakisa (1997) estimated that broken plurals constitute 76%; the remaining 24% form plurals by suffixation.

The claim that Arabic is a minority default system has generated a debate about the structural properties of the morphological model that accounts for the Arabic plural. The main question in this debate is how words in a morphologically complex system like Arabic with regular and irregular inflections are formed and represented. Two models have been proposed to address this question: a rule-based symbolic model and a connectionist network model. Pinker

and Prince (1988) accounts for word formation by proposing a dual-mechanism approach that is equipped with a productive rule-based process to produce regular forms and an associative memory to handle all the irregular patterns. This rule-based model is contrasted with a network connectionist model that uses a memory-based associative network to account for word formation, hence breaking the dependency on rule-based processes. Rumelhart & McClelland (1986) explain that all forms, regular and irregular, in this model are represented in the associative-memory as a network of interconnected units, namely as a connection between input units external to the network and output units with probabilistic contingency, where the strength of connection between input and output units determines the shape and pattern of the form. Proponents of the rule-based symbolic model argue that the status of the Arabic plural as a minority default system makes it impossible for the Arabic plural to be accommodated by a connectionist network model since the condition necessary for establishing a default regular behavior by pluralizing novel input to regular form is absent. The proponents of the connectionist model argue that generalization and productivity is determined, not by symbolic rules, but by the frequency and extent of shared similarity that a form has.

Given the effect of these studies, Arabic plural system was always cited as a prime example of a minority default system. This claim as we will see in the next section was rejected by corpus studies that examined the distribution of sound and broken plural from a usage-based perspective. To the best of our knowledge, all previous support for Arabic as a minority default system has come from dictionary sources, rather than corpus data. Using dictionaries as a data source results in a data-set that is not representative of the language in actual use. For instance, dictionaries are often not up to date for colloquial usage, such as loanwords and lexical

innovations. They may also contain highly specialized words and archaic forms and omit transparently derived regular forms.

### 3.3 Investigating regularity and productivity in the system based on the distribution of sound and broken plurals in previous research

Few studies examined the productivity in the Arabic plural system as implied by the distribution of the nouns between the two plural types. When we search for the studies that counted the occurrences of plural types in language use, the number becomes even smaller. Boudelaa and Gaskell (2002) did an analysis of the statistical distribution of sound and broken plural on 1670 high-frequency nouns from the Basic Lexicon of Modern Standard Arabic (Khouloughli, 1992) with the goal of testing the hypothesis of Arabic having a minority default system of pluralization. In languages with a minority default morphological system, the dominant and productive inflectional process will be one that involves a substantial and complex modification of the input stem (or the irregular marker), whereas the inflectional process that involves little or no allomorphy will apply only to a minority of forms. In Arabic plural system, this would mean the number of nouns that form plurals by taking the non-concatenative broken plural templates will surpass those that take the affixational sound plural. After analyzing the corpus data, the result showed that about 59% of the 1670 most frequent nominal forms pluralize via suffix addition and the remaining forms, around 41%, take a broken plural. Boudelaa and Gaskell then explained that the idea of Arabic having a minority default system of pluralization originates as a result of failure to understand the notion of productivity as a gradient phenomenon, one that distinguishes between qualitative and quantitative productivity. According to them, previous studies that claim Arabic plural is a minority default system ignored the role of qualitative and quantitative productivity on the Arabic plural system. For Boudelaa and Gaskell,

both plural types in Arabic are productive, but in different way. The broken plural is qualitatively productive, whereas sound plural is quantitatively productive.

Dawdy-Hesterberg and Pierrehumbert (2014) examined the frequency distribution of sound and broken plural in 6579 singular-plural pairs collected from the Corpus of Contemporary Arabic (Al-Sulaiti 2009). Consistent with Boudelaa and Gaskell (2002), the percentage of the sound plural in this data (74%) was larger than the percentage of the broken plural. Since the Corpus of Contemporary Arabic provides information about the occurrences of the word type, Dawdy-Hesterberg and Pierrehumbert also examined the number of times each singular-plural pair occurs in the corpus, the token frequency where an instance of a given pair counts as a token. They found higher token frequency for the sound plural (61% word tokens). In sum, the number of nouns that take the affixational sound plural was higher than that of those that take the non-concatenative broken plural, and the nouns with the sound plural are used in running text more frequently than the nouns with the broken plural. Because of these results, Dawdy-Hesterberg and Pierrehumbert agreed with Boudelaa and Gaskell's (2002) that Arabic plural is not a minority default system.

There are numerous studies that investigated the productivity and regularity in the Arabic plural system based on analysis of the statistical distribution of singular stems and their plural forms, yet these analyses were restricted to the nouns that take the broken plural. Levy (1971) and McCarthy and Prince (1990) each did a statistical survey of the most common patterns based on a corpus collected from the dictionary of Wehr (1976). Murtonen (1964) made a comparative study of the statistical productivity of the plural system in Semitic languages by including Arabic data from Lane's (1863) dictionary and Geʕez data from Kazimirski's dictionary. A thorough

review of these studies will be given in the next chapter where we qualitatively analyze broken plurals in Arabic.

### 3.4 Frequency effect on grammar

Over the last three decades, a large amount of research has been focused on examining the effect of frequency on different aspects of grammar. The primary goals of these studies have been 1) to show how speakers' experience represented in language use shape their knowledge of their language, and 2) to propose usage-based models for grammar. One of the prominent usage-based models is the Network model proposed by Bybee (2001). Bybee's (2001) model proposes a strong view of frequency in which morphological and phonological knowledge is viewed as an emergent generalization over the lexicon, where patterns emerge in grammar by means of their lexical strength. The concept of lexical strength of a linguistic form is tied to the notions of frequency and the observation that the regular patterns are those with the highest frequency in the lexicon while the irregular patterns are those with low frequency in the lexicon but high frequency for individual items in natural language use.

Bybee (2001) shows how type and token frequency affect "emergent" generalization over the lexicon. A pattern gleaned from lexical items or words gain strength from the number of unique words sharing it - that is, by their *type frequency*. (Type frequency refers to the single occurrence of a pattern regardless of how frequently it occurs in text.) Bybee indicates that patterns that occur frequently in the lexicon (i.e. those with high type frequency) tend to form a type of regularity or "schemas". Stronger patterns with a large number of participant items are more productive; that is, they are more likely to be used in innovation and regularization. Bybee describes the tendency of patterns with high lexical strength to form regularity and determine productivity as the **type frequency effect**. The other type of frequency effect is the **token**

**frequency effect**. *Token frequency* refers to the frequency of occurrence of a pattern in a corpus of language use. Given the effect of type frequency, less frequent patterns that fade in lexical strength are prone to regularize and align with the stronger pattern. Yet, patterns with low type frequency do exist in the lexicon. Bybee explains that the patterns with low type frequency that occur in the lexicon tend to have high occurrences in actual language use (high token frequency), to make up for their low lexical presence. She suggests that the conserving effect of high token frequency, which protects irregular patterns from regularization on the basis of the productive form, is represented as lexical strength. So, both type and token frequency should contribute to lexical strengths of the pattern.

The usage-based model is influential and offers nuanced perspectives to analyze morphology. Using the analytical tools of the usage-based model, we can extrapolate to predict the status of regular and irregular patterns in a morphological system. First, the usage-based model takes for granted that regular patterns will be the productive process and hence will be more frequent than irregular patterns. Given the effect of the type frequency on productivity, the model predicts that irregular morphological patterns with low type frequency are constantly subject to analogical modeling by regular and productive patterns. However, because token frequency influences the lexical strength of a morphological form, the regularization tendency will be suppressed in irregular patterns with high token frequency. Therefore, in a system where irregular patterns compete with regular productive ones, the model predicts that the irregular patterns have to have high token frequency— they have to be used frequently in order to block the analogical modeling to the regular forms. The Arabic plural system offers an opportunity to test this prediction and examine the merit of the usage-based model.

The results of previous research are inconclusive with regard to Bybee's model predictions. Plunkett & Nakisa (1997) reached a conclusion that contradicts Bybee's prediction about the type frequency as determining regularity. Plunkett and Nakisa estimated that sound plurals, the regular patterns, constitute 24% while Bybee's model entails, in its design, that regular patterns should be the most frequent. Boudelaa and Gaskell (2002) and Dawdy-Hesterberg and Pierrehumbert (2014), on the other hand, agree with the prediction by the usage-based model about the effect of type frequency on regularity. Sound plurals, which represent the regular morphological process, are 59% in Boudelaa and Gaskell (2002) and 74% in Dawdy-Hesterberg and Pierrehumbert (2014). Among previous investigations of the distribution of Arabic nominal plurals, only Dawdy-Hesterberg and Pierrehumbert (2014) reported results about the token count of sound and broken plurals. Their results contradict the usage-based model prediction on the effect of token frequency. While the usage-based model predicts that irregular forms that occur in the system ought to have high token frequency, Dawdy-Hesterberg and Pierrehumbert (2014) found that broken plurals by token occur less frequently than sound plurals.

### 3.5 The current study

In this study, we want to address two main issues in Arabic plural research. First, we want to investigate whether the plural system in Arabic exhibits what is argued by previous studies to be a minority default system. Using corpus-based evidence, we will see if this claim holds true. If Arabic has a minority default system for pluralization, then the broken plural which is supposed to be the irregular pattern will be the default process of pluralization. However, if the hypothesis is not true, the sound plural, which is the regular type of pluralization, will be the dominant process.

The second issue we want investigate in this study has to do with the interaction between the two kinds of frequency as described in the usage-based model of morphology. According to this model, both type and token frequency influence the process of lexical strength of the morphological from upon which regularization and resistance to analogical modeling will be based. Type frequency, on one hand, determines productivity in the system. Token frequency, on the other hand, modulates the malleability or resistance of irregular patterns to regularization by analogical modeling to the most frequent type, the patterns with the high type frequency. Depending on how frequently it occurs in actual language use, irregular patterns may resist or undergo analogical modeling to the most frequent regular pattern. Therefore, the usage-based model predicts that the irregular patterns in the system ought to have high token frequency to resist the regularization to the regular patterns. Patterns that do not occur frequently in actual language use will undergo regularization and analogical modeling to the regular morphological pattern. We want to use the Arabic plural system and the dichotomy between broken and sound plural patterns to test the prediction of the usage-based model. In Arabic plurals, the regular plural formed by affixation competes with about 31 non-concatenative broken plural patterns. If the model's predictions are true, we expect nouns that take the broken plural patterns to have high token frequency. Using a list of singular-plural pairs collected from a corpus of Modern Standard Arabic, we test whether the prediction is true.

Following Boudelaa and Gaskell (2002), we refer to sound and broken pluralization as regular and irregular inflectional processes, respectively. This is not because we think that the former is a rule-based and the latter is not. Rather, we use the term regular as a shorthand for an inflectional process involving little or no allomorphy and the term irregular to describe a process that involves substantial modification of the input.

### 3.6 Data

The data-set consists of (8022) singular-plural pairs collected from a subset of Arabic Gigaword (Parker, Graff, Chen, Kong, & Maeda, 2011), which approximately contains 300,000,000 words. Arabic Gigaword, as the name suggests, is a comprehensive archive of 1,200,000,000 words of newswire text data acquired from Arabic news sources by the Linguistic Data Consortium (LDC) at the University of Pennsylvania. The subset of Arabic Gigaword used in the analysis is news texts collected over a period of 13 years between 1997 and 2010 from three different sources: Al-Ahram, Al-sharq Al-awsat and Al-hayat. The texts include three genres, namely, complete stories, random blurbs, and headlines.

The text in Arabic Gigaword is written in Standard Arabic orthography, which shows the words without the diacritics that mark short vowels, semivowels and geminate consonants. The text is also unannotated or parsed for syntax and morphology. Diacritics and parts of speech (POS) tagging were added by using MADAMIRA (Pasha et al. 2014), which is a software that takes an unannotated Arabic text as an input and performs several tasks, including part of speech tagging, word disambiguation and romanization. When the raw (unannotated) text enters MADAMIRA, a Preprocessor cleans the text and converts it to the Buckwalter representation (Buckwalter 1997), which is a standard approach of conversion of Arabic characters to ASCII characters in a strict one-to-one mapping to perform any upcoming computational analysis. The text is then passed to the Morphological Analysis component, which develops a list of all possible analyses (independent of context) for each word. The text and analyses are then passed to a Feature Modeling component, which applies SVM and language models to derive predictions for the word's morphological features. An Analysis Ranking component then scores each word's analysis list from the Morphological component based on how well each analysis

agrees with the SVM and language model predictions, and then sorts the analyses based on that score. The top-scoring analysis of each word can then be passed to the Tokenization component to generate a tokenization for the word. When all the components have finished, the results are returned to the user, which, in addition to tokenization, include the diacritic forms, lemmas, glosses, morphological features, parts-of-speech.

After POS tagging and diacritization are added, a comprehensive list of plural nouns was compiled. The total count of words in the list is 8726 of potential plural forms. All words in the comprehensive list were checked manually by the author for possible errors and incorrect tagging. Words that had been tagged incorrectly as plural were immediately dropped. Total number of words that have been dropped is 704.

The current analysis was limited to nouns that pluralize either by concatenating the masculine and feminine sound suffixes ([uun] & [aat]) to the singular stem or by taking one of the broken plural patterns. This excludes unsuffixed masculine nouns functioning as collectives, which are common in flora-fauna lexicon, e.g. collective [ʃaɟar] “trees” versus individuative singular [ʃaɟar-at] and individuative plural [ʃaɟar-aat] (used after numerals). It also excludes suppletive plurals like [nisaaʔ] “women” that have no phonologically related singular.

The kind of analyses used in this study requires the Arabic plural data-set to be constructed into pairs of plural forms and their singular stems. To get the singular stems for sound plural, you can remove the sound plural suffix and the remainder will be the singular stem. However, this method does not work with broken plural since the plural formation requires, as the name suggest, breaking the singular stem, and altering its shape by a non-concatenative process (where segments from a stem are inserted in a particular output template). In order to

solve this problem, the lemmas of the words that tagger provided was used as the singular stems for broken plural nouns.

In order to investigate the frequency of plural patterns, broken plural forms have to be converted to CV templates, and sound plural forms are tagged as either masculine or Feminine. To create the templates, broken plurals are entered into a python script that turns each plural noun into a CV sequence, where consonants are coded as C and vowels are coded as V. The results are CV templates for singular and plural forms.

### 3.7 Results

Token and type frequency were counted for each plural type. Results are shown in Figure (3-1). Again, type frequency denotes the number of nouns that take either sound or broken plural and token frequency refers to the number of times each of these types appears in running text. The number of unique nouns or type frequency is roughly equally distributed between the two type of plurals with 4165 nouns taking the broken plural versus 3857 nouns taking the sound plural. However, when we look at their token frequency, the distribution is far from equal, with broken plurals having more token frequency than sound plurals. The nouns that take broken plural patterns constitute 66% of the total tokens that are found in the corpus (N = 3158104).

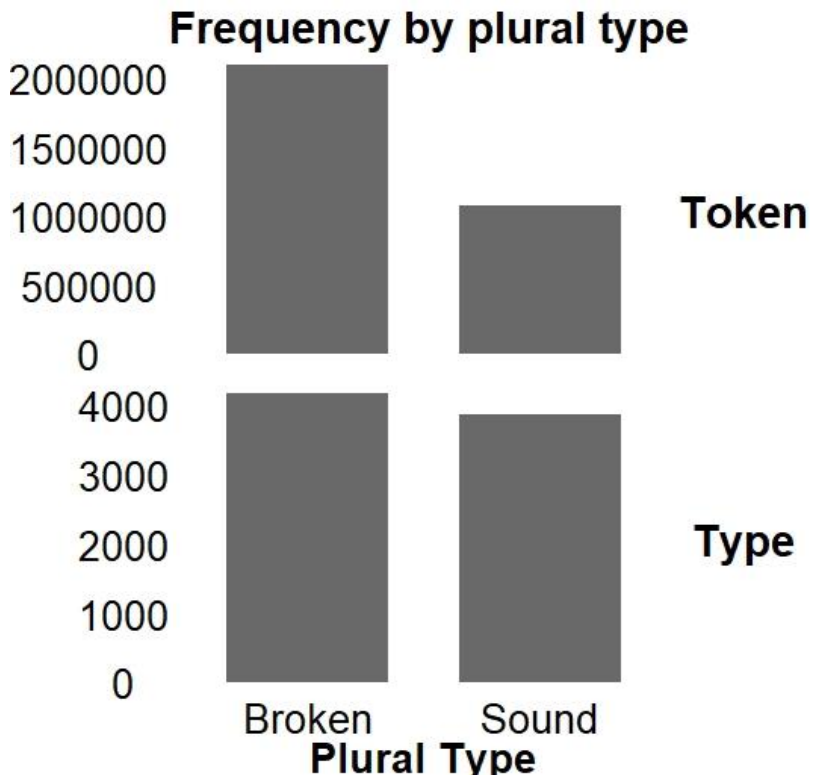


*Figure 3-1.* Type and token counts of sound and broken plural nouns.

In the data, there are 20 CV patterns for the broken plural and 2 suffixes for the sound plural. A breakdown of the noun type (i.e. unique nouns) as distributed between these patterns and suffixes is in Figure (3-2).

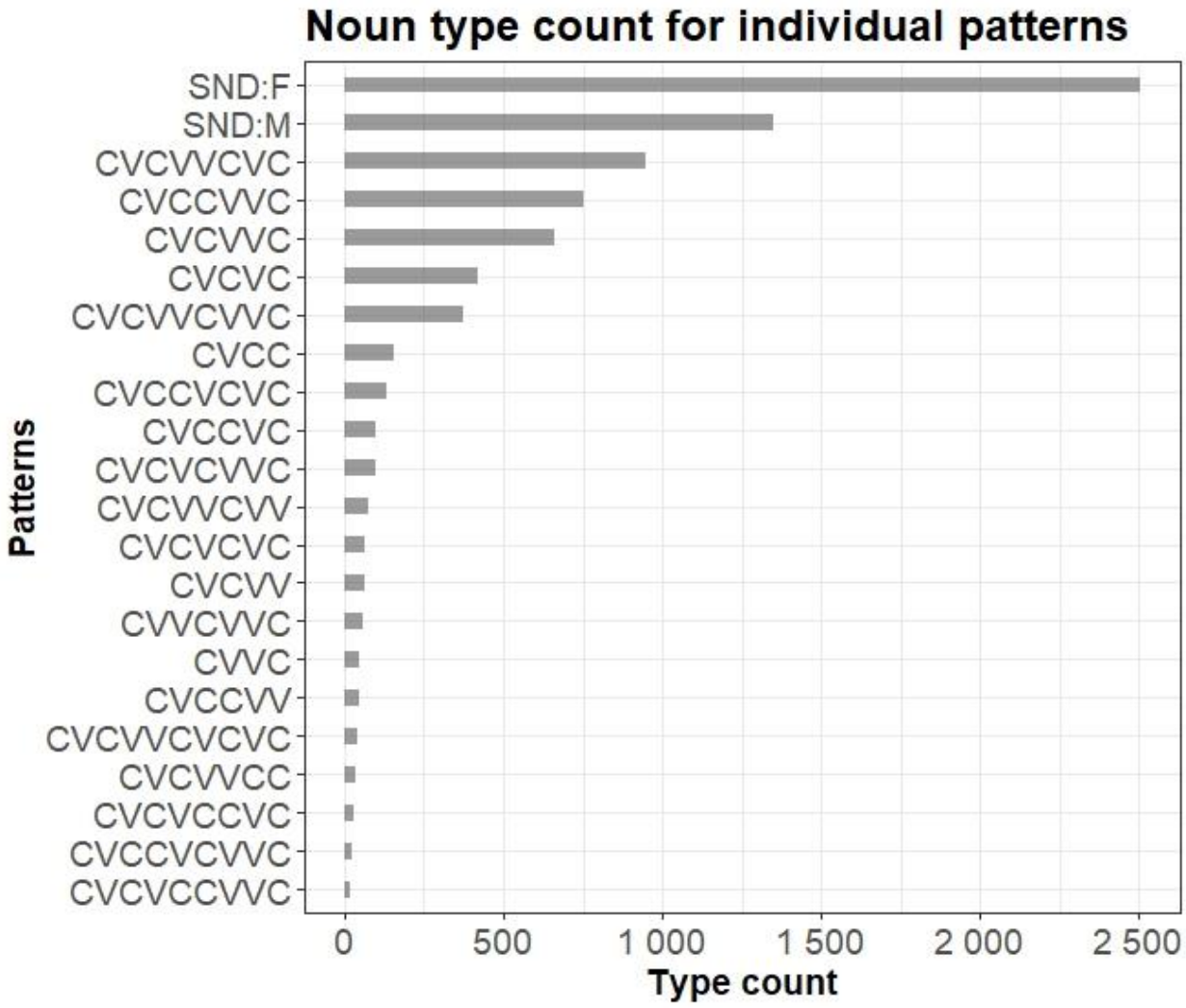


*Figure 3-2.* Type count of the individual patterns within sound and broken plurals.

In binary head to head comparison, the broken plural as single category has more noun types (type count) than the sound plural. However, when each pattern within the sound or broken plural is taken individually, the sound plural has greater type count than the broken plural. Figure (3-2) demonstrates that the number of noun types that take either a feminine (SND:F = 2505) or masculine plural suffix (SND:M = 1352) is greater than the number of noun types in the most frequent broken plural pattern (CVCVVCVC = 947). The distribution of noun types in the broken plural patterns varies dramatically, ranging from 947 to 19 noun types. The token

frequency of nouns that take these patterns was also examined. Figure (3-3) shows the token count of nouns in the individual patterns within sound and broken plural.

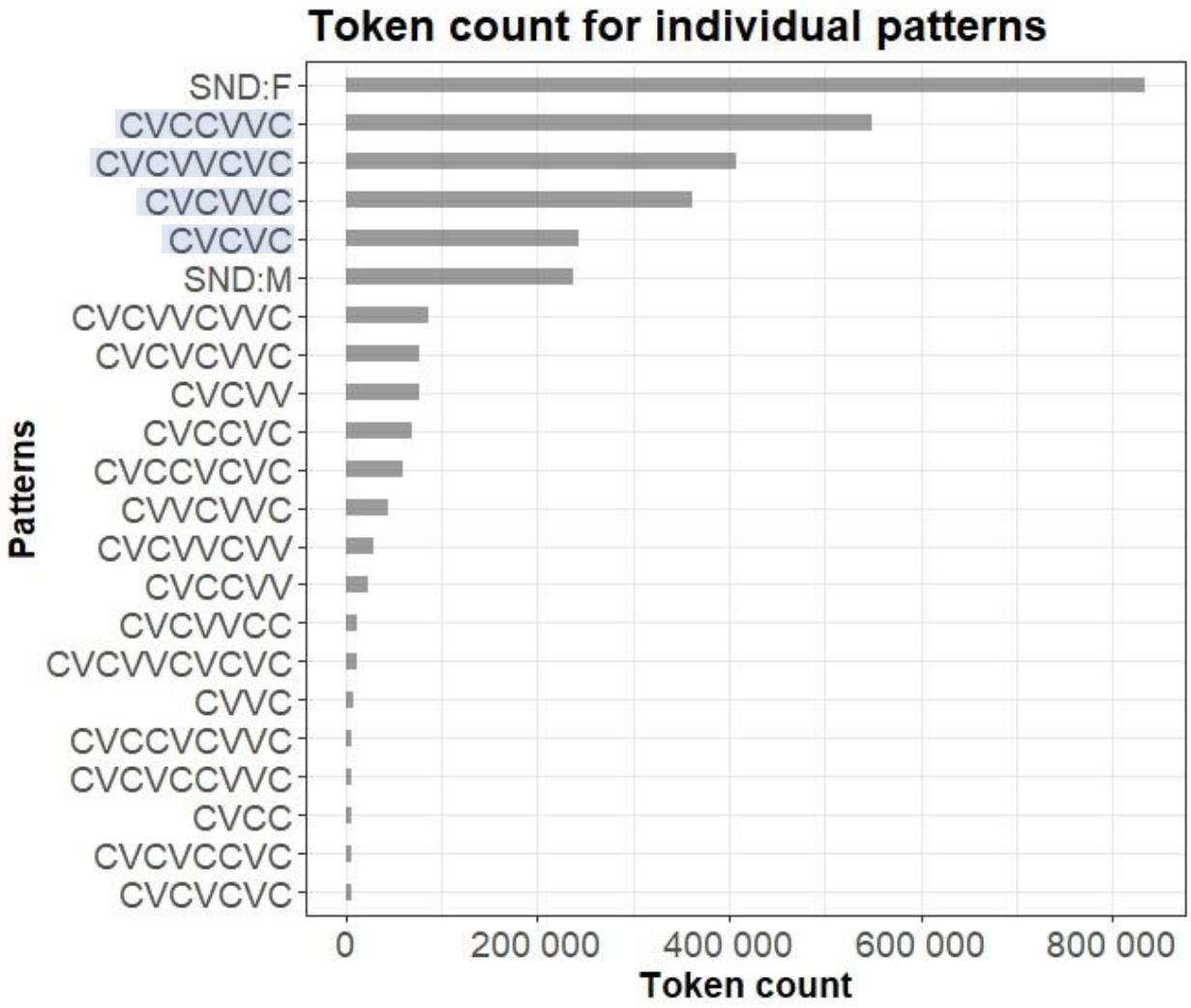


*Figure 3-3.* Token count of nouns in the individual patterns within sound and broken plural.

The statistical analysis applied to the data uses linear regression to model the relationship between plural type (which can be either sound or broken) and the token frequency (the number of times a token that takes one of the plural types occurs in the text). Parametric tests require data that comes from one of the large catalogues of distributions that are thoroughly studied and described by statisticians, and for the data to be analyzed by parametric tests certain assumptions

must be met. If we use a parametric test when the data are not parametric, and the recommended adjustment are not applied then the results are likely to be inaccurate. Therefore, it is important that we check the assumptions before deciding which statistical test is appropriate. Parametric tests require four basic assumptions that must be met for the test to be accurate: a normally distributed sampling distribution, homogeneity of variance, interval or ratio data, and independence.

The assumption of normality of sampling distribution, as the name suggests, checks whether the data comes from a normally distributed sampling distribution. Since it is impossible to have access to the whole sampling distribution from which the data comes, the assessment of the normality assumption is performed on the observed data that we collected. To assess normality visually, a Quantile-Quantile plot for the token frequency scores in the dataset (a continuous dependent variable) is given in Figure (3-4). A Q-Q graph plots the values you would expect to get if the distribution were normal (called here theoretical values) against the values actually seen in the data set (titled sample values). Normality of the distribution is evaluated by showing how the observed values (the dots) fall in the chart. If the data are normally distributed, then the observed values should fall along a straight diagonal line (meaning that the observed values are the same as you would expect to get from a normally distributed data set). Any deviation of the dots from the diagonal line represents a deviation from normality. The scores in the Q-Q plot in Figure (3-4) do not fall in a diagonal line, and the shape of the observed values is representative of a positively skewed distribution. In addition to the visual inspection of normality, I ran an Anderson-Darling normality test to see if the distribution of the scores in the token frequency data significantly deviates from normally distributed scores with the same mean and standard deviation. The result showed that token frequency scores significantly differ from

normally distributed data of the same mean and sd, A = 2030.9, $p$ <.01, and hence the data violates the assumption for normality.

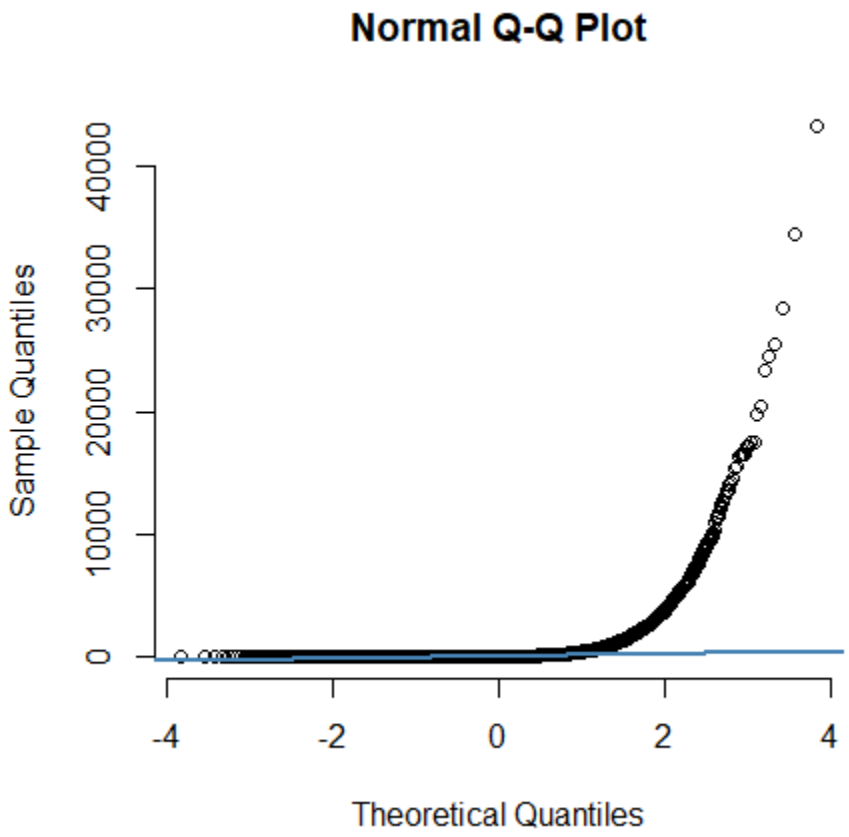


*Figure 3-4.* Quantile-Quantile plot of the token frequency scores.

The second assumption to be assessed is the homogeneity of variance. This assumption requires the variance of the dependent variable to be stable and "homogeneous" at all levels of the independent variable. This means in the dataset of Arabic plural that the variance of token frequency scores should be stable and roughly of the same size in the sound condition as in the broken condition, which are the two levels of the independent variable. Variance that is unstable across the levels of the independent variable represents a violation of this assumption— a heterogeneity of variance. Levene's test was used to examine if the scores had homogeneous variance. It tests the null hypothesis that the variances in different groups are equal (i.e., the

difference between the variances is zero). If Levene's test is significant at $p \leq .05$ then we can conclude that the null hypothesis is incorrect and that the variances are significantly different – therefore, the assumption of homogeneity of variances has been violated. If, however, Levene's test is non-significant (i.e., $p > .05$) then the variances are roughly equal, and the assumption is tenable. For the token frequency scores, Levene's test showed that the variances were significantly different for the sound and broken plural types, $F(1, 8020) = 99.27$, $p<.001$, indicating that the assumption for the homogeneity of variance is not met.

The other two assumptions left are the interval data and the independence of errors. The assumption for interval data is tested by common sense. To say that dependent variable is measured at an interval level, we must be certain that equal intervals on the scale represent equal differences in the property being measured. In the dataset analyzed here, token frequency is measured by counting the number of times one of its instances occurs in running text, such that each time the token occurs in text, its token frequency increases by one value. For this scale to be at interval level, it must be the case that the difference between a word with 20 token frequency and another with 30 token frequency is the same as the difference between a word with 60 and another with 70 token frequency. Since this is case in token frequency scores, I can say that the dependent variable in this study is measured at the interval level. The assumption for independent errors requires the errors in the regression model to be uncorrelated. This assumption will be tested after running the regression model.

Investigation of the assumptions for parametric tests revealed that the data violates two assumptions: the normality of sampling distribution and the homogeneity of variance. A standard practice to correct for problems of non-normality and unequal variance is to transform the data (i.e. token frequency scores). The appropriate transformation method to pursue when the data is

diagnosed with positive skewness and unequal variance is log-transformation which transforms the scores in data to their natural logarithms. Taking the logarithm of a set of scores squashes the right tail of its distribution. As such it is a good way to reduce positive skewness and make the distribution of scores resemble a normal distribution that is needed for the predictions of the statistical model to be accurate. Hence, instead of running the linear regression on the untransformed scores, the linear regression analysis will be done on the log of the token frequency.

Once the token frequency scores were log-transformed, we ran a simple linear regression model in which the log of token frequency scores were entered as outcomes (the values that will be predicted), and the plural type was entered as a predictor. The results of the regression model are given in Table (3-1) and show that there was a significant log-linear relationship between token frequency and plural types, $F(1, 8020) = 165.1$, $p<.001$, $R^2 = .02$, such that as the predictor variable changes from broken to sound plural type the log of token frequency is predicted to decrease (by 0.68 score), $B = .71$, $p<.001$. (Standardized coefficients are not reported because the predictor is categorical. Standardization of the predictor variable is necessary to interpret the change in the outcome when the predictor is continuous because it frames the expected change in the outcome as a result of the change in predictor in units we can understand, thus making the interpretation of the regression model easier. However, categorical variables should not be standardized or centered because when the predictor is categorical, the standardization does not help or simplify the interpretation. So there is no need to report the standardized beta to interpret the change in the predictor that is categorical as in the case of plural type).

| | ***R*²** | ***B*** | ***SE B*** | ***P*** |
|---|---|---|---|---|
| | 0.02 | | | |
| **Constant** | | 3.56 | 0.04 | <.001 |
| **Plural Type** | | -0.68 | 0.05 | <.001 |

*Table 3-1.* Results of the regression model.

To visualize the relationship between the outcome and the predictor, the result of the regression model was plotted in Figure (3-5). In this figure, plural types are plotted in the x-axis while the log of token frequency is plotted in the y-axis. The upper and the lower bounds of the 95% confidence intervals are computed and added for the two plural types. A confidence interval is a range of scores calculated such that the population mean will fall within this range in 95% of samples. Therefore, comparison of the confidence intervals of two means gives a chance to see if these means come from two different populations, so that they are significantly different. If the intervals of two means does not overlap, we can infer the means are from different populations whereas if there is an overlap, this suggest the difference between the means are not significant. As indicated by the results in Table (3-1), Figure (3-5) shows a decrease in token frequency as the plural type changes from broken to sound. Also, the Figure shows no overlap between the error bars representing the 95% confidence interval for the means of each plural type. Therefore, we can infer that the two means are from different population - they are significantly different.

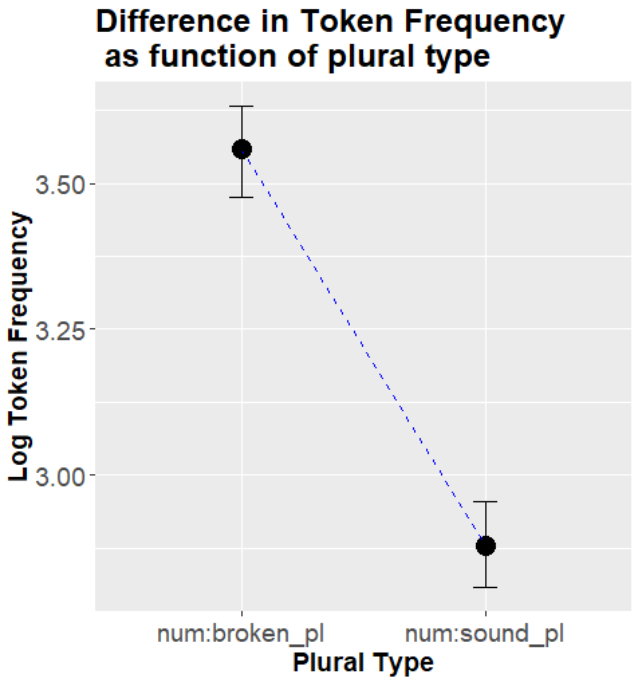


*Figure 3-5.* The results from the regression model showing the decrease in the Log Token Frequency as a result of the change from broken plural (num:broken_pl) to the sound plural (num:sound_pl). Error bars represent the 95% confidence intervals for each plural type.

## 3.8 Discussion

The results showed that the noun types that take the sound plural (48%) are slightly fewer than those that take the broken plural (Figure 3-1). This finding is inconsistent with the results from Boudelaa and Gaskell (2002) and Dawdy-Hesterberg and Pierrehumbert (2014), which argued that the Arabic plural is not a minority default system. However, the difference between sound and broken by type count is marginal as the two plural types differ by less than 2%.

The results contradict the prediction by the usage-based model concerning the effect of type frequency on determining the regular pattern. The usage-based model predicts that the majority of forms will take the regular inflection. The results, on the contrary, show that the number of noun types that pluralize via the regular sound suffixes are fewer than the number of noun types that take broken plural patterns (Figure 3-1).

I acknowledge that the conclusions about the type frequency effect are based on the type count of nouns in the broken and sound plurals when each plural type is treated as a whole category. If the type frequency of the individual patterns within these plurals is considered, the results might lead to a different conclusion that is not necessarily inconsistent with the prediction by the usage-based model. The results of the individual plural patterns within sound and broken plural demonstrate that the type count for the most frequent broken pattern is smaller than that for the masculine or feminine sound inflection (Figure 3-2), indicating that pluralization via one of the sound suffixes is the most frequent inflection in Arabic.

This research also shows that token frequency which represents the number of times a plural noun appears in actual language use differs significantly between the two types of plurals. This does not mean that there is a negative correlation between type and token frequency such that an increase in type frequency of forms results in a decrease in its token frequency because this is inconsistent with research on the type-token vocabulary curve (Youmans 1990) which shows that as the count of unique forms type increases, their token frequency increases too. However, what the results in the current study show is that although the number of nouns that take sound plurals is roughly equal to the number of nouns that take broken plurals, on average the broken plural is used more often than the sound plural.

This pattern of relationship between token frequency and the irregular morphological inflections is also partly reflected in the token frequency of individual patterns within the sound and broken plurals (Figure 3-3). When the plural patterns are investigated individually, four of the top five most frequent patterns (CVCCVVC, CVCVVCVC, CVCVVC, and CVCVC) belong to the broken plural.

The significance of this result lies in the insights that that Arabic plural system can give on understanding the relationship between regularity and productivity. In the Arabic plural system, the difference between the two ways of making plural, in fact, represents a difference between a productive regular plural and a less productive irregular type. According to the usage-based model of morphology, type frequency of regular morphological patterns, indicated by the number of words that take the regular pattern, tends to be higher than that of the irregular patterns. Because the regular pattern is more productive than the irregular complex patterns, there is always a tendency in language for the irregular minority to be regularized or changed by analogical modeling to the most frequent lexically strong patterns. However, the irregular patterns that do not regularize tend to occur with high token frequency. This increase in use by the irregular patterns is argued to block the tendency toward regularization (Bybee 2001). This prediction is supported by the results from this study, which shows that irregular patterns are used more frequently than the regular ones even when their type frequency is roughly equal.

As the results demonstrate, the data of token frequency is positively skewed as a result of a large number of the data being packed on the small end of the token frequency scale, and also has unequal variance across the two plural types. These issues are not trivial and usually indicate some hidden structure in the data that current factor (independent variable) failed to unveil. These issues may be accounted for by certain variables that were not included in the current analysis. One of the factors that may explain these issues, but the current analysis unfortunately did not address is related to semantic characteristics and their influence on repression (decreasing) or encouraging (increasing) the use of nouns. Finding this hidden structure and solving the problems of skewing and variation in the data is a problem for future research to address.

Lack of factors that could account for the variance in the data is one of the limitations in the current study. Another limitation is related to language varieties and the genres from which the data is collected. The analysis draws conclusions collected from a corpus of written text but does not include data from a spoken variety. Furthermore, the text from which the plural-singular pairs are extracted and analyzed covers three genres: political news, sports, and book blurbs. There is a possibility that less diverse data with a small number of genres could result in biased and outcomes and inaccurate analysis. A recommendation for future research would be to use data that covers more genres and is representative of actual language use in its spoken and written forms.

# Chapter 4 Investigation of the Role of Stem Weight on the Formulation of Broken Plural Nouns in Arabic

## 4.1 Introduction

In this chapter, I will discuss the role of the weight of the stem on the mapping of singular stems to their plural form. Weight is the way of describing and analyzing morphemes by measuring a certain quantity. Previous research has shown that weight plays a role on several phonological and morphological phenomena (e.g. Clements & Keyser 1983, Hyman 1985, Newman, 2017). The broken pluralization in Arabic is an interesting case to examine the effect of the segmental weight of a stem input on the selection of a plural form.

## 4.2 Review of the previous analyses of Arabic plural system

Numerous attempts have been made to account for Arabic broken plurals. These accounts fall into three main groups according to their specific morphological approach: Generative Morphophonology models, Root-&-Pattern Morphology, and Prosodic Morphology. This section will give a review of each approach.

### 4.2.1 Rule-based transformational models within the generative framework

Most of the earlier studies adopt a generative rule-based approach to Arabic plural form derivation, e.g., Brame (1970) and Levy (1971). According to these studies, Arabic plural forms are derived via a series of complex, abstract rules that must apply in a certain order. This approach assumes that forms have a surface as well as underlying representation, and the process of generating forms involves transformation of the underlying form by means of reorganizing the

limited set of phonological rules in grammar. The following will be a review of two studies that use this approach to account for Arabic plural form derivation.

Brame (1970) did a general analysis of Arabic morphology from a generative rule-based perspective. For Arabic broken plural, the general argument in this analysis is that singular nouns that take the same plural pattern should share a common form at the underlying level. Brame, for example, cites the examples in Table 4-1 below, where the singular stems in (a) have the shape CvCCvC, whereas the singular stems in (b) have the shape CvCvvC, yet they take the same broken plural template:

| | Singular | Plural | Gloss |
|---|---|---|---|
| a. | maktab – at<br>ʔarmal - at | makaatib<br>ʔaraamil | "library"<br>"widow" |
| b. | sˁaħiif - at<br>dˁariib - at | sˁaħaaʔif<br>dˁaraaʔib | "newspaper"<br>"tax" |

*Table 4-1.* Singular stems in (a) and (b) have different shapes, yet they take the same broken plural template.

Brame claims that stems in (b) and (a), regardless of the difference in their syllabic shape, at the underlying level have the same *CvCCvC* shape. According to Brame, the long vowel [ii] in the final syllable of stems in (b) is surfacing from an underlying sequence of a glide and a high vowel /ji/. So, the underlying forms of [sˁaħiif] and [dˁariib] become /sˁaħjif/ and /dˁarjib/. Brame lists the following rules to derive the surface representations of the stems in (b):

Glide metathesis: GvC -> vGC
Vowel assimilation: j -> i / i_C. also w -> u / u_C

These rules are applied in the following order to generate the surface forms from their hypothetical underlying representations:

| | | |
|---|---|---|
| Underlying: | /sˁaħjif/ | /dˁarjib/ |
| Glide metathesis: | sˁaħijf | dˁarijb |
| Vowel assimilation: | sˁaħiif | dˁariib |
| Surface: | [sˁaħiif] | [dˁariib] |

Brame proceeds to argue that the plural forms for stems in (b) are also derived by applying the rule of plural formation, which is insertion of /aa/ after the second consonant of the singular stem, to the underlying representations of these stems rather than their surface forms. During the derivation of plural from the underlying forms, the underlying form is subject to a glide formation rule which changes the glide to a glottal stop:

Glide formation: G -> ʔ / aa__

The derivation of the plural forms of the stems in (b) from their underlying representations is given below:

| | | |
|---|---|---|
| Underlying: | /sˤaħjif/ | /dˤarjib/ |
| Plural rule: | sˤaħaajif | dˤaraajib |
| Glottal formation: | sˤaħaaʔif | dˤaraaʔib |
| surface: | [sˤaħaaʔif ] | [dˤaraaʔib] |

Levy (1971) provided a similar account and applied the same transformational rule-based model but with an expanded set of data and more emphasis on the derivation of broken plural nouns. The data in her study includes pairs of singular and plural nouns collected from Wehr's (1976) dictionary. Beside the account for the singular to plural mapping, Levy also provides a statistical analysis and information about the frequency distribution of the singular stems of their plural templates.

### 4.2.2 Root-&-Pattern model

One of the earliest lessons that students in introductory linguistic courses are taught is that words in Semitic languages are uniquely formed by combining a consonantal root, which indicates a core meaning, and a syllabic pattern, which indicates a grammatical function, (e.g. Coles 1996; Watson 2002). This approach was then responsible for a series of studies that lay the groundwork for what came to be known as the Root-&-Pattern theory. Under this type of analysis, a broken plural form is derived by inserting the consonantal root from the singular stem

onto a CV template. Two of the most notable studies within the framework of Root-&-Pattern model are McCarthy (1979) and Hammond (1988).

Influenced by the insights of Goldsmith's (1976) Autosegmental Phonology, McCarthy proposes a model that accounts for word formation in a nonlinear morphological system like Arabic by subdividing the word form into three elements: a consonantal root, a vowel series, and a syllabic template. Although McCarthy's model takes for granted the notions of root and pattern, his innovation is in the split of the traditional pattern into a vocalic melody and CV template. Thus, the representation of the plural form of the noun in his model must include a consonantal tier, which provides the root, a vowel tier, which provides the vocalic melody, and a CV skeleton in the middle. For example, McCarthy presents the following scheme to show how the plural form [ɟanaadib] 'locusts' is derived:

a-

ɟundub ɟanaadib

b-

Quadriliteral noun redundancy rule:

[*CvCCv(v)C*] $_{\text{singular}}$ ~ [*CvCvvCv(v)C* ]$_{\text{plural}}$

c-

Root tier:

[ɟndb]

Template

[*CvCvvCv(v)C*]

vowel tier:

a – i

The scheme shows the derivation of the plural [ɟanaadib] for the singular [ɟundub]. First, the rule in (a) gives the prosodic templates (for singular and plural) for quadriliteral nouns. According to the rule in (a), the plural template associated with the quadriliteral singular [ɟundub] is [*CvCvvCv(v)C*]. Derivation then proceeds by mapping radical consonants from the root tier to their corresponding positions in the plural CV template. Mapping of the consonants to

their C positions in the plural template will strictly follow Goldsmith's (1976) Universal Convention of left-to-right association creating the filled template *ɟvnvvdv(v)b*. The vowels from the vocalic melody tier are then mapped to their positions in the plural template according to a language specific rule. The second vowel from the plural vocalic melody is mapped to the V position in the final space of the plural template to generate [ɟvnvvdi(v)b]. Given that the vowel in the second syllable of the singular stem is short, McCarthy states /i/ will fill one vowel position and the vowel between parenthesis will be dropped. The remaining vowel /a/ from the vocalic melody goes to fill the empty V position in the first and the second syllable yielding [ɟanaadib].

Hammond's (1988) account, on the other hand, is actually a critique of McCarthy's account and offers some alternative proposals. The only major difference between these two studies is in the mechanism by which the consonants are mapped onto the plural template. Hammond proposes that instead of the direct linking between the consonantal root and the template, consonants are associated to the plural template via transfer from the singular stem to the plural template to account for some cases that otherwise remain unexplained in McCarthy's original analysis.

### 4.2.3 Prosodic Morphology

In an attempt to provide a comprehensive analysis for the Arabic broken plural, McCarthy and Prince (1990) propose a theory of Prosodic Domain Circumscription that draws heavily on the interface between phonology and morphology by establishing a link between the templates and the prosodic structure. They argue that "rules sensitive to morphological domain may be restricted to a prosodically characterized (sub-)domain in a word or stem" (McCarthy and Prince, 1990:209). According to McCarthy and Prince, Prosodic Domain Circumscription

relies on the principles of the theory of Prosodic Morphology developed in McCarthy and Prince (1986). These central principles are:

I. Prosodic Morphology Hypothesis. Templates are defined in terms of the authentic units of prosody: mora ($\mu$), syllable ($\sigma$), foot (F), prosodic word (W), and so on.

II. Template Satisfaction Condition. Satisfaction of templatic constraints is obligatory and is determined by the principles of prosody, both universal and language-specific.

III. Prosodic Circumscription of Domains. The domain to which morphological operations apply may be circumscribed by prosodic criteria as well as by the more familiar morphological ones. In particular, the minimal word within a domain may be selected as the locus of morphological transformation in lieu of the whole domain.

(McCarthy and Prince 1990:209)

McCarthy and Prince use the patterns of broken plural listed in Wright (1971), after dividing them into four prosodically defined categories. The four categories can be seen in Table (2-2) in chapter 2, which is reproduced below for convenience:

| **a. Iambic** | **b. Trochaic** | **c. Monosyllabic** | **d. Other** |
|---|---|---|---|
| *CiCaaC* | *CuCaC* | *CuCC* | *CuC$_j$C$_j$aC* |
| *CuCuuC* | *CiCaC* | *CiCC + at* | *CuC$_j$C$_j$aaC* |
| *CaCaaC* | *CaCaC* | *CiCC + aan* | |
| */CaCaaC/* surfacing as *ʔaCCaaC* | *CiCaC + at* | *CuCC + aan* | |
| *CaCaaC + /ay/* | */CaCuC/* surfacing as *ʔaCCuC* | *CaCC + /ay/* | |
| *CaCiiC* | *CuCuC* | *CaCC* | |
| *CuCuuC + at* | *CaCaC + at* | | |
| *CiCaaC + at* | *CuCaC + at* | | |
| *CawaaCiC* | *CuCaC + aaʔ* | | |
| *CaCaaʔiC* | */CaCiC/ + at* surfacing as *ʔaCCiCat* | | |
| *CaCaaCiC* | */CaCiC/ + aaʔ* surfacing as *ʔaCCiCaaʔ* | | |
| *CaCaaCiiC* | | | |

*Table 2-2*. The four groups of most frequent plural patterns from McCarthy and Prince based on Wright's.

The four categories in Table (2-2) are labeled according to their prosodic characteristics: all the forms in (a) begin with the iambic foot *CvCvv+;* the forms in (b) are all quantitative trochee *CvCvC* (i.e., a bimoraic foot with a final extrasyllabic consonant); and all the forms in (c) are monosyllabic *CvCC.*

McCarthy and Prince argue that these four classes are not of equal importance. In particular, the monosyllabic plural patterns in (c) and the plural patterns labeled 'other' in (d) are of limited interest. According to McCarthy and Prince, although the monosyllabic plural patterns are widespread among different singular stem shapes, they occur at very low frequencies, whereas the plural patterns *Cu*$C_jC_j$*a(a)C*, labeled 'other', are narrowly restricted to singular stems of the shape *CaaCiC* which are lexicalized active participles.

The trochaic patterns in (b) are accounted for by a root-and-pattern approach along the lines of McCarthy and Prince (1986), where the consonants of the root are mapped onto a disyllabic quantitative trochee template, consisting of two moras and two syllables, with a final extrasyllabic consonant: *CvCv<C>*.

McCarthy and Prince develop the Prosodic Domain Circumscription analysis mainly to account for the iambic broken plural patterns in (a) above. They argue that despite the fact that the iambic patterns are the dominant class in the lexicon, they cannot be represented adequately using the familiar resources of the Root-and-Pattern approach. Unlike the trochaic patterns, the derivation of iambic plural patterns follows from their characterization in prosodic terms. Their approach to derivation of the iambic broken plurals from the singular stems can be summed up in three steps. First, the first two moras in the singular are isolated (circumscribed) so that a minimal word is created. Second, the minimal word is then mapped onto an iambic foot (*CvCvv* template). Third, the remainder of the singular stem that is left after the circumscription is

attached to the end of the plural form and the vowel melody is changed. An example showing how these steps are executed is provided below:

| | | |
|---|---|---|
| *[sultaan]* | *[nafs]* | ***Singular stem*** |
| *sul* | *naf* | ***Isolate a minimal word (two moras)*** |
| *suluu* | *nafaa* | ***Map onto an iambic template*** |
| *salaatiin* | *nufuus* | ***Attach the remainder and change vowels*** |

McCarthy and Prince's work on Arabic broken plural in terms of the Prosodic Domain Circumscription has evolved to become an influential line of research. Prosodic Morphology provides powerful analytical tools that were not available in the previous models to account for challenging morphological problems. Insights from Prosodic Morphology and its application on Arabic Plural, for example, have contributed to significant advances in phonological theory, and specifically on the framework of Optimality Theory (Prince and Smolensky 1993).

### 4.3 Weight as a phenomenon that influences phonology

Linguists have discussed the concept of weight and quantity in two ways. It has been associated with different elements based on the areas of research. In prosodic phonology, heaviness is used as a property of the syllable, which is classified as heavy or light based on metrical criteria. In the field of morphology, the heavy and light are used to describe stems, and a stem is said to be heavy or light on the basis of quantitative approach such as the number of its segments. In what follows, I describe and show the differences between each of these methods.

#### 4.3.1 Metrical methods to represent syllable weight

Two representations of weight that have gained wide acceptance in phonological theory are the skeletal slot models (McCarthy 1979; Clements & Keyser 1983), and moraic models (Hyman 1985; Hayes 1989). The appeal of the two models is that they both argue for the need of representations that express the mutual independence of segmental quantity (represented in

segment count and phonemic duration) and segmental quality. Both models describe the quantity of syllables on the basis of the weight units assigned to the segments within the syllables, but they differ in their definition of the unit that can bear weight. In the skeletal tier model, Clements and Keyser (1983) propose that segments are not immediately associated to the syllables but are dominated by timing structure known as skeletal slots, which encode (both) segment duration and syllable weight. Consonants and vowels that are linked to single slots are short, while those that have long vowels and geminate consonants are linked to double slots. They also assume there are two types of slots, *C*, where *C* represents syllable margins (coda or onset), and *V*, which represents syllable nucleus. Each of these timing slots also acts as a weight unit, where a syllable is described as heavy or light depending on the number of slots it receives. A syllable that is linked to two timing slots in a form of *CV* is said to be light, whereas three timing slots in a type of *CVV* or *CVC* describe a heavy syllable.

The skeletal tier models have received criticism for its limitation to account for the difference in status between segments in the rhyme and segments in the onset. Hayes (1989) listed two cases where the difference is highlighted. The first case is the compensatory lengthening, where Hayes (1989) pointed that compensatory lengthening always occurs in the case of segments deleted from the rhyme, and never occurs in the case of segments deleted from the onset. Hayes also noted that the difference between segments in the rhyme and those in the onset is revealed by their contribution to syllable weight and quantity, where quantity is determined mainly by the number of segments in the rhyme, again to the exclusion of the onset. This suggests that segments in the rhyme have a special status when it comes to syllable weight.

In order to capture the difference in status between segments in the rhyme and segments in the onset, alternative models adopt the mora as the basic unit of weight (Hyman 1995; Hayes

1989). In lieu of the skeletal tier, the mora becomes the intermediate level that links segments with prosodically active status to the syllable. Segments in the syllable rhyme are prosodically active, hence, they are given moras, while consonants in the onset, which are said to be prosodically inactive, do not receive moras but are directly linked to the syllable node. A language with no vowel-length distinction and that does not allow a coda will only have syllables with one mora. A language which has a vowel-length distinction will have both monomoraic and bimoraic syllables. The weight distinction for syllables then becomes a distinction between monomoraic syllables which encode light weight and bimoraic syllables that encode heavy weight.

### 4.3.2 Additive approach to describe stem weight

In addition to the skeletal slot and moraic models, there has been a call for an alternative approach that does not rely on metrical structure to determine weight (e.g. Heath 2005, 2018). The resort to this method of representing weight is presented as a solution to describe morphology distinctions displayed by stems that cannot be contained in conventional metrical approaches. Unlike the previous analysis where weight is measured by breaking words into syllables and then counting moras or CV positions in the syllables within a word, the current approach measures weight by directly adding the number of segments in the word. So, the weight in this approach is not a feature of the syllables but rather the whole stem. At the intellectual level, the two methods seem to represent two different schools of thought. The metrical models follow to a bottom-up tradition where weight is measured by breaking words into syllables where CV positions or mora are counted (weight units are contained within syllables which are contained within words), whereas the additive approach adopts a top-down process where weight is considered a property for the whole word, not its constituent. Below I

review some of the morphological problems that do not lend themselves to the prosodic units, but rather the distinction is based on the number of the segments within the word.

One of the numerous examples that show a morphological distinction caused by stem weight is in tonal patterning in verbs in Penange. In Penange, the effect of stem weight on tonal patterns can be seen in many inflectional categories of the verb. In these categories, the tonal overlay in the verbs exhibits a distinction between two classes of verb stems: light stems (which can take any of these forms *Cv*, *Cv:*, *CvNCv* or *CvCv*) and heavy verbs (which take the following patterns *CvCCv* with *CC* not homorganic cluster, *Cv:Cv* or *CvCvCv*) (Heath 2018:10). Verb inflectional categories where the exact weight split is displayed include the imperatives and 3rd singular perfective verbs (henceforth 3Sg). Examples illustrating this split in the imperative in Penange is given below:

| | 3Sg perfective | Imperative | Gloss |
|---|---|---|---|
| Light stem | | | |
| | nɛ́ | ná | say |
| | nɛ́: | ná: | drink |
| | sígé | sígó | go down |
| Heavy stem | | | |
| | bá:ndè | bà:ndà | shut door |
| | nó:yè | nò:yè | sleep |
| | yígírè | yìgìrò | shake |

*Table 4-2.* Tonal patterning in Penange, from Heath (2018:187).

In Table (4-2) above, High tone (H) overlays in all light 3Sg perfective and imperative verbs while low tone (L) in all 3Sg perfective and imperative verbs that belong to heavy weight. Such a problem could not be described using conventional metrical models.

In Tamashek verbal morphology, long and short imperfective of verbs are formed by overlaying the stems with one of three vocalic melodies: low <L>, High <H>, or High Low <HL> (Heath 2005:324). Stems with lexical u and most with lexical i (except Ci…) have <H>

vocalism. For stems with only short vowels “v” representing ə and æ, the overlay depends on weight:

| | stem | short Ipfv | long Ipfv | vocalism |
|---|---|---|---|---|
| Light | vCCvC | əPPəC | t-áPPæC | H / L |
| | vPQvC | əPQəC | PáQQæC | H / L |
| | vPvC | əPəC | (t-)əPPáC | H / (H)L |
| Heavy | PvCvC | æPPæCæC | t-áPəCɑC | L / L |
| Superheavy | PvQvCCvC | əPQəCCəC- | t-íPQəCCiC | H / H |
| stems with only short vowels “v” representing ə and æ, overlay depends on weight: | | | | |

*Table 4-3.* Tamashek verbal morphology, from Heath (2005:324).

Another phenomenon where such stem weight affect morpho-phonological processes is in the area of templatic fit. English and Arabic both use segmental weight of the input to determine templatic fit. In English, this is illustrated in comparative -er and diminutive -y where input stems fit into a template or are rejected depending on their size. For comparative -er and diminutive -y, the template accepts only light inputs (unless truncated) (e.g. comparative: slow >> slower; heavy >> heavier), while heavy inputs are rejected (*interestinger, *capabler). The difference between light and heavy stem in the selection of the template for comparative and diminutive is demonstrated in (4-4):

| A. English -er | |
|---|---|
| big >> bigger<br>heavy >> heavier<br>slow >> slower | interesting >> *interestinger<br>beautiful >> *beautifuler<br>supportive >> *supportiver |
| B. English -y | |
| John >> Johny<br>dog >> doggy<br>pig >> piggy | Joseph >> *Josephy<br>rabbit >> *rabbity |

*Table 4-4.* English comparative and diminutive.

The use of segmental weight to gauge template fit also occurs in Arabic. Rather than using weight to reject inputs as in English comparative and diminutive, weight in Arabic functions as triage to sort input into classes and select template variants that will be adapted to fit

each particular input class. Examples of such morphological distinction that is based on input weight include the derivation of active and passive participles from perfective verbs. Derivation of new words from inputs involve mapping the segments from an input stem to a template. Yet the shape of that template will be adapted depending on the number of segments in the input stem, or stem weight. Examples to illustrate this morphological process are given in (4-5):

| **Weight** | **Adapt template for active participle** | **Adapt template for passive participle** | |
|---|---|---|---|
| A. **Light input** | CaaCiC | ma-CCuuC | gloss |
| katab | kaatib | maktuub | "writer" |
| ʃarib | ʃaarib | maʃruub | "drink" |
| **B. Heavy input** | mu-…CiC | ma-…CaC | |
| daħraɟ | mudaħriɟ | mudaħraɟ | "roll" |
| ɟaħfal | muɟaħafil | muɟaħafal | "beat" |
| baʃʃar | mubaʃʃir | mubaʃʃar | "announce" |

*Table 4-5.* Active and passive participles in Arabic

The table in (4-5) shows how segmental weight of the stem input affects derivation of active and passive participles in Arabic. After the weight of stem inputs is evaluated, the template is adapted to fit classes of inputs. Triconsonantal light verbs form active and passive participles with the templates *CaaCiC* and *maCCuuC* while heavy verbs (those with four consonants, or in some cases with three consonants and a long vowel) take the template mu-*…CiC* for active participle and *mu-…CaC* for passive. Once template variants for each class of input are created, new words can be produced by mapping the segments from inputs to their corresponding positions in the outputs. For example, mapping segments from the perfective verb katab to its active and passive participle templates will produce kaatib and maktuub, whereas the input segments from the verb daħraɟ will produce mudaħriɟ for active participle and mudaħraɟ for passive participle.

Previous examples showed that languages are capable of employing quantitative weight of the input to sort inputs and modify output templates in order to derive new forms in a top-down fashion. It seems that, among the several attempt to analyze broken plural in Arabic, there has not been any investigation of the influence of the stem weight on the derivation of the plural form. In the next section I try to analyze Arabic nominal plural by focusing on the role of the weight of stem inputs on the production of plural forms.

## 4.4 Data

Unique plural-singular pairs of broken plurals used in the analysis of statistical distribution of Arabic plural patterns in chapter 3 are used to analyze the broken plural system in Arabic. The data have 4165 singular-plural pairs of noun stems. To get the vocalized patterns required for the analysis, the syllabic and vocalic structures are taken from every singular and plural form. Extracting all the vocalized patterns from the data revealed that there are 94 singular stem patterns mapped to 43 plural patterns.

## 4.5 Analysis

Plural patterns with more than 15 forms are shown in Table (4-6). In the table, the left column is for the plural patterns and the column to its right is for patterns of singular stems that are mapped to plural patterns in the left. At the right of every singular form there are four numbers. The first number is the count of singular stems that have this pattern out of all singulars taken by a given plural pattern. The second is the count of singulars with this pattern regardless of plural form. The third number is the percentage of singulars with this pattern in relation to all singulars mapped to a given plural pattern. The fourth number is percentage of a plural pattern in relation to all plural patterns taken by a given singular pattern. For example, the row for the plural *ʔaCCiCaaʔ* has the following singular *CaCiiC* and next to the singular stem there are these

numbers [21, 337, 1.0, 0.06]. These numbers indicate that 21 of the singulars mapped to *ʔaCCiCaaʔ* have the pattern *CaCiiC*, 337 of all singulars take the pattern *CaCiiC*, 100 percent of the singulars mapped to plural *ʔaCCiCaaʔ* have the pattern *CaCiiC*, and only 6 percent of all singulars *CaCiiC* have a plural *ʔaCCiCaaʔ*.

| Plural | Singular stems | Count |
|---|---|---|
| CaCaaCaa | 'CaCCaaC': [20, 68, 0.27, 0.29], 'CiCCaC': [1, 150, 0.01, 0.01], 'CaCiC': [1, 24, 0.01, 0.04], 'CiCaaCaC': [1, 21, 0.01, 0.05], 'CaCC': [1, 546, 0.01, 0.0], 'CiCCaaC': [2, 66, 0.03, 0.03], 'CuCCaaC': [1, 36, 0.01, 0.03], 'CaaCiCaC': [1, 114, 0.01, 0.01], 'CuCCaa': [3, 7, 0.04, 0.43], 'CaCCaaCiC': [1, 1, 0.01, 1.0], 'CaCiiCaC': [3, 127, 0.04, 0.02], 'CaCiiC': [4, 337, 0.05, 0.01], 'CaCiCCaC': [29, 30, 0.4, 0.97], 'CaCCaa': [4, 54, 0.05, 0.07], 'CaCCaC': [1, 412, 0.01, 0.0] | 73 |
| ʔaCCuC | 'CaCaaC': [4, 57, 0.07, 0.07], 'CiCCaC': [1, 150, 0.02, 0.01], 'CuCC': [1, 107, 0.02, 0.01], 'CuCaaC': [5, 22, 0.09, 0.23], 'CaCaC': [2, 240, 0.04, 0.01], 'CiCaaC': [5, 111, 0.09, 0.05], 'CiCC': [2, 150, 0.04, 0.01], 'CaCC': [32, 546, 0.58, 0.06], 'CaCiiC': [3, 337, 0.05, 0.01] | 55 |
| CiCCaan | 'CaCaaC': [1, 57, 0.05, 0.02], 'CaCaa': [1, 13, 0.05, 0.08], 'CuCaaC': [3, 22, 0.14, 0.14], 'CaCiC': [2, 24, 0.1, 0.08], 'CaCaC': [4, 240, 0.19, 0.02], 'CaCC': [3, 546, 0.14, 0.01], 'CaCiiC': [3, 337, 0.14, 0.01], 'CaaCC': [3, 75, 0.14, 0.04], 'CuCaC': [1, 1, 0.05, 1.0] | 21 |
| CuCaat | 'CiCCiC': [2, 9, 0.03, 0.22], 'CaCiiC': [1, 337, 0.01, 0.0], 'CaaCC': [67, 75, 0.93, 0.89], 'CuCCaC': [1, 232, 0.01, 0.0], 'CaCCaC': [1, 412, 0.01, 0.0] | 72 |
| ʔaCCiCat | 'CaCaaC': [34, 57, 0.25, 0.6], 'CiCCaC': [1, 150, 0.01, 0.01], 'CuCaaC': [9, 22, 0.07, 0.41], 'CaCiC': [2, 24, 0.01, 0.08], 'CaCaC': [2, 240, 0.01, 0.01], 'CiCaaC': [56, 111, 0.41, 0.5], 'CiCC': [1, 150, 0.01, 0.01], 'CaCC': [3, 546, 0.02, 0.01], 'CaCiiCaC': [1, 127, 0.01, 0.01], 'CaCiiC': [16, 337, 0.12, 0.05], 'CaCuuC': [5, 27, 0.04, 0.19], 'CaaCC': [4, 75, 0.03, 0.05], 'CiCiiC': [1, 1, 0.01, 1.0] | 135 |
| ʔaCiCCat | 'CiCCiiC': [1, 25, 0.03, 0.04], 'CuCaaC': [3, 22, 0.1, 0.14], 'CiCaaC': [7, 111, 0.24, 0.06], 'CaCiiC': [16, 337, 0.55, 0.05], 'CaCuuC': [2, 27, 0.07, 0.07] | 29 |
| CuCCaaC | {'Total': 160, 'CuCC': [1, 107, 0.01, 0.01], 'CuCaaC': [1, 22, 0.01, 0.05], 'CaCaC': [8, 240, 0.05, 0.03], 'CiCaaC': [2, 111, 0.01, 0.02], 'CaCC': [5, 546, 0.03, 0.01], 'CaCiiC': [20, 337, 0.12, 0.06], 'CaCaCC': [1, 28, 0.01, 0.04], 'CaCCaa': [1, 54, 0.01, 0.02], 'CaCCaC': [4, 412, 0.03, 0.01], 'CaaCiC': [117, 333, 0.73, 0.35]} | 160 |
| CiCaC | 'CiiCaC': [14, 16, 0.13, 0.88], 'CiCCaC': [80, 150, 0.73, 0.53], 'CaCaCaC': [1, 11, 0.01, 0.09], 'CiiCaaC': [1, 13, 0.01, 0.08], 'CaCCaC': [13, 412, 0.12, 0.03] | 109 |
| CuCuC | 'CaCaaC': [5, 57, 0.05, 0.09], 'CaCCaaC': [3, 68, 0.03, 0.04], 'CiCaaCiC': [1, 1, 0.01, 1.0], 'CuCC': [1, 107, 0.01, 0.01], 'CaCiC': [4, 24, 0.04, 0.17], 'CaCaC': [4, 240, 0.04, 0.02], 'CiCaaC': [31, 111, 0.3, 0.28], 'CaCaCaC': [1, 11, 0.01, 0.09], 'CaCaaCaC': [1, 15, 0.01, 0.07], 'CaCC': [1, 546, 0.01, 0.0], 'CaCiiCaC': [7, 127, 0.07, 0.06], 'CaCiiC': [27, 337, 0.26, 0.08], 'CaCuuC': [12, 27, 0.12, 0.44], 'CuCCaC': [1, 232, 0.01, 0.0], 'CaaCiC': [4, 333, 0.04, 0.01] | 103 |
| CaCaaCiC | 'CiiCaC': [1, 16, 0.0, 0.06], 'CaCaaC': [5, 57, 0.01, 0.09], 'CaCCiCaC': [18, 18, 0.02, 1.0], 'CuCCuCaC': [5, 5, 0.01, 1.0], 'CaCCaaC': [6, 68, 0.01, 0.09], 'CaCCiCCaC': [13, 13, 0.01, 1.0], 'CiCCiC': [6, 9, 0.01, 0.67], 'CuuCiC': [1, 2, 0.0, 0.5], 'CiiCiCCaC': [1, 1, 0.0, 1.0], 'CaCCaCiC': [1, 5, 0.0, 0.2], 'CuCCiCaC': [2, 2, 0.0, 1.0], 'CiCCaC': [48, 150, 0.05, 0.32], 'CuuCaa': [1, 1, 0.0, 1.0], 'CiCCaa': [1, 1, 0.0, 1.0], 'CaCCuC': [1, 1, 0.0, 1.0], 'CuCaaC': [1, 22, 0.0, 0.05], 'CaCCaCaC': [75, 76, 0.08, 0.99], 'CiCaaC': [3, 111, 0.0, 0.03], 'CuCaaCaC': [1, 1, 0.0, 1.0], 'CiCC': [1, 150, 0.0, 0.01], 'CuCCuC': [26, 28, 0.03, 0.93], 'CiCaaCaC': [20, 21, 0.02, 0.95], 'CaCaaCaC': [14, 15, 0.01, 0.93], 'CaCC': [5, 546, 0.01, 0.01], 'CiCCaaC': [9, 66, 0.01, 0.14], 'CaaCiCaC': [113, 114, 0.12, 0.99], 'CaCCaaCaC': [1, 9, 0.0, 0.11], 'CaaCaC': [6, 10, 0.01, 0.6], 'CuCaCCaaC': [1, 1, 0.0, 1.0], 'CaCCaCuuC': [1, 3, 0.0, 0.33], 'CaCuC': [1, 2, 0.0, 0.5], 'CiiCaaC': [2, 13, 0.0, 0.15], 'CaCuuCaC': [3, 3, 0.0, 1.0], 'CaCiiCaC': [116, 127, 0.12, 0.91], 'CaCiiC': [25, 337, 0.03, 0.07], 'CuCCiC': [9, 15, 0.01, 0.6], 'CaCCaCiiC': [1, 2, 0.0, 0.5], 'CiCCiiCaC': [3, 3, 0.0, 1.0], 'CuCiiCaC': [2, 2, 0.0, 1.0], 'CiCCaCaC': [30, 30, 0.03, 1.0], 'CiCCiCaC': [3, 3, 0.0, 1.0], 'CaCuuC': [2, 27, 0.0, 0.07], 'CuCCaC': [7, 232, 0.01, 0.03], 'CaCCiC': [37, 43, 0.04, 0.86], 'CaCCaa': [49, 54, 0.05, 0.91], 'CaCCaC': [172, 412, 0.18, 0.42], 'CuCCiCCaC': [14, 14, 0.01, 1.0], 'CaaCiC': [82, 333, 0.09, 0.25], 'CuCaCCaC': [2, 2, 0.0, 1.0] | 947 |
| ʔaCiCCaaʔ | 'CaCCaaC': [1, 68, 0.05, 0.01], 'CaCiiC': [18, 337, 0.95, 0.05] | 19 |
| ʔaCCaaC | 'CiiCaC': [1, 16, 0.0, 0.06], 'CiCCiiC': [1, 25, 0.0, 0.04], 'CaCaaC': [3, 57, 0.0, 0.05], 'CaCaa': [3, 13, 0.0, 0.23], 'CiCCaC': [1, 150, 0.0, 0.01], 'CaaC': [9, 27, 0.01, 0.33], 'CuuC': [18, 26, 0.03, 0.69], 'CuCC': [80, 107, 0.13, 0.75], 'CiiC': [16, 25, 0.03, 0.64], 'CaCiC': [3, 24, 0.0, 0.12], 'CiCaC': [1, 1, 0.0, 1.0], 'CaCaC': [201, 240, 0.33, 0.84], 'CiCaaC': [3, 111, 0.0, 0.03], 'CaCaCaC': [2, 11, 0.0, 0.18], 'CiCC': [90, 150, 0.15, 0.6], 'CaCC': [150, 546, 0.25, 0.27], 'CuCuC': [5, 5, 0.01, 1.0], 'CaCiiC': [11, 337, 0.02, 0.03], 'CuCCiC': [2, 15, 0.0, 0.13], 'CuCCiiC': [1, 5, 0.0, 0.2], 'CaCuuC': [1, 27, 0.0, 0.04], 'CuCCaC': [1, 232, 0.0, 0.0], 'CaCCiC': [3, 43, 0.0, 0.07], 'CaaCiC': [2, 333, 0.0, 0.01] | 608 |
| CuCaCaaʔ | {'Total': 100, 'CaCaaC': [1, 57, 0.01, 0.02], 'CuCiiC': [1, 1, 0.01, 1.0], 'CaCiiC': [90, 337, 0.9, 0.27], 'CaaCiC': [8, 333, 0.08, 0.02]} | 100 |
| CuCaC | 'CuuCaC': [20, 20, 0.11, 1.0], 'CuCCaC': [155, 232, 0.83, 0.67], 'CaCCaC': [11, 412, 0.06, 0.03] | 186 |

| | | |
|---|---|---|
| CiCaaC | 'CaCaaC': [1, 57, 0.0, 0.02], 'CaCCaaC': [5, 68, 0.02, 0.07], 'CiCCaC': [2, 150, 0.01, 0.01], 'CaaC': [3, 27, 0.01, 0.11], 'CuCC': [6, 107, 0.03, 0.06], 'CiiC': [2, 25, 0.01, 0.08], 'CaCiC': [5, 24, 0.02, 0.21], 'CaCaC': [13, 240, 0.06, 0.05], 'CiCaaC': [1, 111, 0.0, 0.01], 'CaCaCaC': [7, 11, 0.03, 0.64], 'CiCC': [11, 150, 0.05, 0.07], 'CaCC': [58, 546, 0.25, 0.11], 'CaaCaC': [1, 10, 0.0, 0.1], 'CaCuC': [1, 2, 0.0, 0.5], 'CaCiiC': [46, 337, 0.2, 0.14], 'CaCuuC': [5, 27, 0.02, 0.19], 'CuCCaC': [16, 232, 0.07, 0.07], 'CaCCaC': [42, 412, 0.18, 0.1], 'CaaCiC': [4, 333, 0.02, 0.01] | 229 |
| CuCCaC | 'CaCiC': [1, 24, 0.03, 0.04], 'CaCC': [1, 546, 0.03, 0.0], 'CaCCaC': [1, 412, 0.03, 0.0], 'CaaCiC': [30, 333, 0.91, 0.09] | 33 |
| CaCaCat | 'CaCiiC': [1, 337, 0.02, 0.0], 'CaaCiC': [57, 333, 0.98, 0.17] | 58 |
| CuCuuC | 'CiCCiC': [1, 9, 0.0, 0.11], 'CiCCaC': [1, 150, 0.0, 0.01], 'CuCC': [17, 107, 0.05, 0.16], 'CiiC': [3, 25, 0.01, 0.12], 'CaCiC': [3, 24, 0.01, 0.12], 'CaCaC': [6, 240, 0.02, 0.03], 'CiCC': [39, 150, 0.11, 0.26], 'CaCC': [261, 546, 0.75, 0.48], 'CaCiiC': [1, 337, 0.0, 0.0], 'CuCCiC': [2, 15, 0.01, 0.13], 'CuCCaC': [1, 232, 0.0, 0.0], 'CaCCaC': [7, 412, 0.02, 0.02], 'CaaCiC': [6, 333, 0.02, 0.02] | 348 |
| CiiCaan | 'CaaC': [10, 27, 0.43, 0.37], 'CuuC': [8, 26, 0.35, 0.31], 'CiiC': [1, 25, 0.04, 0.04], 'CaCC': [3, 546, 0.13, 0.01], 'CaaCiC': [1, 333, 0.04, 0.0]} | 23 |
| CaCaaCiiC | 'CiCCiiC': [19, 25, 0.05, 0.76], 'CaCCuuC': [67, 67, 0.18, 1.0], 'CaCCaaC': [11, 68, 0.03, 0.16], 'CuCCuuCaC': [14, 14, 0.04, 1.0], 'CaCCaCaC': [1, 76, 0.0, 0.01], 'CaCCiiCaC': [11, 11, 0.03, 1.0], 'CaaCuuC': [34, 34, 0.09, 1.0], 'CaaCiiC': [2, 2, 0.01, 1.0], 'CaCCuuCaC': [10, 10, 0.03, 1.0], 'CaCC': [1, 546, 0.0, 0.0], 'CiCCaaC': [50, 66, 0.13, 0.76], 'CuCCaaC': [30, 36, 0.08, 0.83], 'CaCCaaCaC': [8, 9, 0.02, 0.89], 'CaaCuuCaC': [10, 10, 0.03, 1.0], 'CiiCaaC': [10, 13, 0.03, 0.77], 'CiCCaCC': [3, 4, 0.01, 0.75], 'CaCiiC': [1, 337, 0.0, 0.0], 'CaCCaCiiC': [1, 2, 0.0, 0.5], 'CuCCiiC': [4, 5, 0.01, 0.8], 'CuuCaaC': [1, 1, 0.0, 1.0], 'CaCCiC': [1, 43, 0.0, 0.02], 'CiCCaaCaC': [2, 2, 0.01, 1.0], 'CaCCiiC': [43, 43, 0.11, 1.0], 'CuCCuuC': [39, 40, 0.1, 0.97], 'CaaCiC': [1, 333, 0.0, 0.0] | 374 |
| CaCCaa | 'CaCCaaC': [8, 68, 0.18, 0.12], 'CaCiC': [2, 24, 0.05, 0.08], 'CaCC': [4, 546, 0.09, 0.01], 'CaCiiC': [26, 337, 0.59, 0.08], 'CaCCiC': [2, 43, 0.05, 0.05], 'CaaCiC': [2, 333, 0.05, 0.01] | 44 |
| CaCaaCiCat | 'CiCCiiC': [4, 25, 0.1, 0.16], 'CaCaaC': [2, 57, 0.05, 0.04], 'CaCCaaC': [2, 68, 0.05, 0.03], 'CuuCiC': [1, 2, 0.02, 0.5], 'CaCCaCiC': [4, 5, 0.1, 0.8], 'CaCCiCaaC': [1, 1, 0.02, 1.0], 'CaaCuuCiC': [1, 1, 0.02, 1.0], 'CuCCuC': [2, 28, 0.05, 0.07], 'CaCCiCiC': [1, 1, 0.02, 1.0], 'CiCCaaC': [5, 66, 0.12, 0.08], 'CuCCuCiC': [1, 1, 0.02, 1.0], 'CuCCaaC': [5, 36, 0.12, 0.14], 'CaCCuuCiC': [2, 2, 0.05, 1.0], 'CaCCaCuuC': [2, 3, 0.05, 0.67], 'CiCCaCC': [1, 4, 0.02, 0.25], 'CuCCiC': [1, 15, 0.02, 0.07], 'CaCCiiCiC': [1, 1, 0.02, 1.0], 'CaCCaC': [4, 412, 0.1, 0.01], 'CuCCuuC': [1, 40, 0.02, 0.03] | 41 |
| CuuC | 'CaaC': [3, 27, 0.07, 0.11], 'CiCaaC': [2, 111, 0.04, 0.02], 'CaaCaC': [3, 10, 0.07, 0.3], 'CaCCaC': [38, 412, 0.83, 0.09] | 46 |
| CaCaaCC | 'CaCaCCaC': [1, 1, 0.03, 1.0], 'CaCaCCaaaC': [22, 22, 0.63, 1.0], 'CaaCCaC': [10, 10, 0.29, 1.0], 'CaCiCCaC': [1, 30, 0.03, 0.03], 'CaCaCC': [1, 28, 0.03, 0.04] | 35 |
| CuCC | 'CaCCaaC': [3, 68, 0.02, 0.04], 'CaCC': [6, 546, 0.04, 0.01], 'CuCCaa': [2, 7, 0.01, 0.29], 'CaCiiC': [1, 337, 0.01, 0.0], 'CuCCiC': [1, 15, 0.01, 0.07], 'CaCaCC': [26, 28, 0.18, 0.93], 'CaCCaC': [107, 412, 0.72, 0.26], 'CaaCiC': [2, 333, 0.01, 0.01] | 148 |
| ʔaCCiCaaʔ | 'CaCiiC': [21, 337, 1.0, 0.06] | 21 |
| CuCaa | 'CiCCaC': [1, 150, 0.02, 0.01], 'CuCCaa': [2, 7, 0.04, 0.29], 'CuCCaC': [48, 232, 0.89, 0.21], 'CaCCaC': [3, 412, 0.06, 0.01] | 54 |

*Table 4-6.* Plural patterns and their singular stems.

I can notice several observations from the table in (4-6). Every plural template is linked with multiple singular patterns. This indicates the relationship between plural forms and singular stem inputs can be described as a one-to-many relationship. The table also shows the possibility of grouping plural forms by the shape of the singular stem input. The main criterion that defines a set of singular stems that dominantly take given plural forms is the number of segments (consonants and full vowels) within these stems. Stems that are linked to a plural template may not have the exact syllabic shape or vocalism, but they certainly contain in one way or the other the same quantity of consonants and full vowels. As we saw in the previous section, this agreement or specification on the quantity of segments by a set of stem input that take the same

output form is a well-known phenomenon in variety of unrelated languages. It is, also, another way to represent the weight or heaviness of morphological forms.

Depending on the weight of the singular stem, the plural forms are classed into three major weight groups: light, middle and heavy. The light group includes stems with the three or fewer consonants. This class of stems take the following plural forms: *CuCuuC*, *ʔaCCaaC*, *ʔaaCaaC*, *ʔaCCuC*, CiCaaC, *CiCCaan*, *CiiCaaC* and *CiCaCat*. Stems with four or more consonants belong to the heavy class which is mapped to the following plurals: *CaCaaCiiC*, *CaCaaCiC*, *CaCaaCaa*. The remaining class of middle weight covers a special group of stems that have four segments most often in the form of three consonants and one long vowel.

### 4.5.1 Light stems

**CvC(v)C > CuCuuC, ʔaCCaaC, ʔaCCuC, CiCaaC, CiCCaan, ʔaaCaaC**

Stems with three consonants and no long vowel are almost predominantly associated with one of the six patterns above. Some nouns have two or more plural forms from this set. Examples of each plural pattern is given in Table (4-7):

| | | | | |
|---|---|---|---|---|
| CaCC | | | | |
| qafr | qifaar | | | “wasteland” |
| sabʕ | sibaaʕ | subuuʕ | | “predator” |
| kalb | kilaab | | | “dog” |
| baħr | biħaar | buħuur | ʔabħaar | “sea” |
| ɟaħʃ | ɟiħʃaan | ɟuħuuʃ | | “donkey” |
| kahl | kuhuul | | | “elderly person” |
| θawb | θiyaab | ʔaθwaab | | “dress” |
| ʔalf | ʔaalaaf | ʔuluuf | | “thousand” |
| ʔany | ʔaanaaʔ | | | “moment” |
| ʔalw | ʔaalaaʔ | | | “blessing” |
| raʔy | ʔaaraaʔ | | | “opinion” |
| CuCC | | | | |
| ɟurħ | ɟuruuħ | ɟiraaħ | | “wounds” |
| qurtˁ | qiraatˁ | ʔaqraatˁ | | “earrings” |
| ɣusˁn | ɣusˁuun | ʔaɣsˁaan | | “branch” |
| CiCC | | | | |

| | | | |
|---|---|---|---|
| qidr | quduur | | “pot” |
| ʕiɟl | ʕuɟuul | | “calf” |
| sˁibɣ | sˁibaaɣ | ʔasˁbaaɣ | “paint” |
| biʔr | ʔaabaar | | “well” |
| CaCaC | | | |
| ʕalaf | ʔaʕlaaf | ʕilaaf | “fodder” |
| balad | bilaad | | “country” |
| θamar | ʔaθmaar | θimaar | “crop” |
| ðakar | ðukuur | | “male” |
| waral | ʔawraal | wirlaan | “lizard” |
| walad | ʔawlaad | wildaan | “boy” |
| ħadaθ | ʔaħdaaθ | ħidθaan | “young” |
| ʔadab | ʔaadaab | | “literature" |
| ʔaθar | ʔaaθaar | | “fossil” |
| CaCuC | (This is the only example) | | |
| raɟul | riɟaal | | “man” |
| CaCiC | | | |
| raħim | ʔarħaam | | “womb” |
| malik | muluuk | | “king” |
| kabid | kubuud | | “liver” |

*Table 4-7.* Light weight stems.

In the table above, we can notice a remarkable similarity between *ʔaCCaaC* and *ʔaaCaaC*. The only difference in the syllabic structure of the two patterns is in their first syllable which can be either a coda-less syllable (*ʔaa*) or a closed one (*ʔaC*). A possible explanation for the similarity is that the pattern *ʔaaCaaC* is derived from *ʔaCCaaC* but has undergone a phonological change that led it to surface as *ʔaaCaaC*. In terms of preference by their singular stems, the two patterns seem to be in complementary distribution. Plurals with *ʔaaCaaC* are restricted to singular stems with an initial or medial glottal stop. (I must say that triconsonantal stems with a glottal in the second *C* are, in fact, rare, to the extent that [raʔy] and [biʔr] are the only two stems that exist in my data). The *ʔaCCaaC* pattern, on the other hand, never associates with singular stems with an initial glottal. However, when the segments from the singular input are mapped to their positions in the plural template, the template position where the glottal stop is expected is filled by a low vowel /a/. Thus, initial glottal stops in the singular stem forms a

plural by taking a *ʔaCCaaC* template, then anticipatory assimilation to the preceding low vowel leads the glottal stop to surface as vowel in the template *ʔaaCaaC*. Thus, we can argue that *ʔaaCaaC* is derived from *ʔaCCaaC*.

A small number of light singular stems that are mapped to one pattern from this group also have another pattern from outside the group. All the light stems that take plurals that do not belong to the patterns (*CuCuuC*, *ʔaCCaaC*, *ʔaCCuC*, *CiCaaC*, *CiCCaan*) are listed below in (4-8):

| light stem | inside group | outside group | |
|---|---|---|---|
| ʕabd | ʕibaad | ʕabiid | “worshiper” |
| kahl | kuhuul | kuhhal | “elderly person” |
| ʔamr | ʔumuur | ʔawaamir | “command” |
| θaman | ʔaθmaan | ʔaθminat | “value” |
| waθan | ʔawθaan | wuθun | “idol” |
| ʃarak | ʔaʃraak | ʃuruk | “trap” |
| (No examples for ʔaCCuC) | | | |

*Table 4-8.* Light stems that take plurals that do not belong to the patterns (CuCuuC, ʔaCCaaC, ʔaCCuC, CiCaaC, CiCCaan).

It is not uncommon for a plural to have singulars of different types, but the distribution of these singular patterns is far from random. Usually a plural form will be strongly associated with one or two singular patterns, and the rest of the patterns will be infrequent. The distribution of the most frequent singular types, as indicated by percentage, for each of these six plural forms is illustrated in Table (4-9) below:

| Output | Input | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| CiCaaC | CaCC (26%) | CaCiiC (21%) | CaCCaC (18%) | CuCCaC (7%) | CaCaC (6%) | CiCC (5%) | CuCC (3%) | CaCaCaC (3%) | CaCiC (2%) |
| CuCuuC | CaCC (74%) | CiCC (11%) | CuCC (5%) | CaCCaC (2%) | CaCaC (2%) | CaaCiC (2%) | | | |
| ʔaCCaaC | CaCaC (29%) | CaCC (26%) | CiCC (15%) | CuCC (14%) | CuuC (3%) | CiiC (3%) | CaaC (2%) | CaCiiC (2%) | |
| ʔaCCuC | CaCC (60%) | CuCaaC (7%) | CiCaaC (7%) | CaCaaC (7%) | CiCC (5%) | CaCaC (5%) | CaCiiC (5%) | CuCC (2%) | |
| CiCCaan | CaCaC (21%) | CaCC (16%) | CaCiiC (16%) | CuCaaC (16%) | CaCiC (11%) | CuCaC (5%) | | | |
| ʔaaCaaC | CaCaC (40%) | CaCC (20%) | CiCC (16%) | CuCuC (8%) | CaCaCaC (4%) | CaCiiC (4%) | | | |

*Table 4-9.* Most frequent singulars for the six plural patterns in the light-weight group.

Based on the percentage, the basic singulars that are most frequently mapped to these six plural forms are *CaCC*, *CaCiiC*, *CaCaC*, *CiCC*, *CuCC*. The percentage of the six plurals as taken by these basic singular nouns is as follows:

| | CiCaaC | CuCuuC | ʔaCCaaC | ʔaCCuC | CiCCaan | ʔaaCaaC | Collective percentage |
|---|---|---|---|---|---|---|---|
| CaCC | 11% | 47% | 28% | 5% | 1% | 1% | 93% |
| CaCiiC | 16% | 0 | 4% | 1% | 1% | 0 | 22% |
| CaCaC | 6% | 3% | 75% | 7% | 2% | 5% | 98% |
| CiCC | 8% | 26% | 57% | 1% | 0 | 3% | 95% |
| CuCC | 6% | 17% | 72% | 1% | 0 | 0 | 96% |

*Table 4-10.* The percentage of the light-weight plurals as taken by the basic singular stems.

The proportion of individual singulars as distributed between these plurals indicates some type of preference in the selection of plural patterns. All five plural patterns *CuCuuC*, *CiCaaC*, *ʔaCCaaC*, and *ʔaCCuC* are favored by singulars of the type *CaCC*, but the largest proportion of these singulars are mapped to *CuCuuC* and *ʔaCCaaC*. Singulars with *CaCaC*, *CuCC* and, to a lesser degree, *CiCC* show preference for plurals with pattern *ʔaCCaaC*. The second most preferred form for singulars with patterns *CiCC* and CuCC is *CuCuuC*. The third most preferred plural for the high vowel light stems is *CiCaaC*. Overall, the sum of percentage of each singular type exceeds 90 % for the four singular patterns *CaCC*, *CaCaC*, *CiCC* and *CuCC*. The exception is *CaCiiC* which reaches a total of 22 %. Out of the five plural patterns, the preferred one for singular stems of *CaCiiC* type is *CiCaaC*. Almost all of the singular stems with the pattern *CaCiiC* are adjectives that describe size. Thus, the plural *CiCaaC* seem to form a special type of plural for these singular adjectives. Some examples are given in Table (4-11) below:

| Singular | Plural | |
|---|---|---|
| kabiir | kibaar | “large” |
| sˁaɣiir | sˁiɣaar | “small” |
| samiin | simaan | “bulky” |
| qasˁiir | qisˁaar | “short” |
| tˁawiil | tˁiwaal | “tall” |

*Table 4-11.* Singular stems of the shape CaCiiC.

Vowel quality seemingly influences the process of plural formation from a group of light stem singulars. Plurals of the type *CuCuuC* and *ʔaCCuC* which have a high vowel (/u/) are associated with singular stems with low vowel *CaCC* and *CaCaC*. Also, singulars with high vowel such as *CiCC* and *CuCC* show preference for plurals with pattern *ʔaCCaaC* which has a low vowel. However, the vowel polarity argument seems to fall apart once we consider the preferences reflected by other singulars stems and the distribution of other plural patterns. The same low vowel plural *ʔaCCaaC* which seemed to be favored by singulars with high vowels is also preferred by stems with low vowel *CaCaC*. The argument for vowel quality polarity influence on plural formation does not seem to hold for the plural patterns *CiCaaC* and *CiCCaan*.

#### 4.5.1.1 Biconsonantal stems with a long vowel

Biconsonantal stems with a long medial vowel form the plural by exclusively taking one of these patterns: *ʔaCCaaC*, *CiCaaC*, *CiiCaan*, *CuCuuC* and *CiCaCat*.

| | | | | |
|---|---|---|---|---|
| CaaC | | | | |
| maal | ʔamwaal | | | “money” |
| baab | ʔabwaab | biibaan | | “door” |
| naab | ʔanjaab | nujuub (only example) | | “canine” |
| daar | dijaar | dijarat (only example) | | “home” |
| ɟaar | ɟiiraan | | | “neighbor” |
| CaaC stems never take a plural of CaaCaaC | | | | |
| CiiC | | | | |
| diik | dijakat | dujuuk | | “rooster” |
| tˁiib | tˁujuub | ʔatˁjaab | | “perfume” |
| fiil | fijalat | fujuul | ʔafjaal | “elephant |
| riiħ | rijaaħ | ʔarjaaħ | | “wind” |
| CiiC stems never take a plural of CiiCaan type | | | | |
| CiiC stems never take a plural of CaaCaaC type | | | | |
| CuuC | | | | |
| ruuħ | ʔarwaaħ | | | “spirit |
| ʕuud | ʔaʕwaad | ʕiidaan | | “stick” |
| kuuʕ | ʔakwaaʕ | kiiʕaan | | “elbow” |

| | | |
|---|---|---|
| suuq | ʔaswaaq | “market” |
| duud | diidaan | “worm” |
| CuuC stems never take a plural of CaaCaaC type | | |
| CuuC stems never take a plural of CiCaCat type | | |
| CuuC stems never take a plural of CiCaaC type | | |
| CuuC stems never take a plural of CuCuuC type | | |

*Table 4-12*. Biconsonantal stems with a long vowel.

Long vowel biconsonantal stems are infrequent and form a minority group. So, looking at their percentage or count in comparison to the other singular stems for a particular plural would not be very informative. A more informative approach is to analyze the distribution of each singular type between the five plural patterns shown above. As indicated by percentage, the three biconsonantal stems *CaaC*, *CiiC*, *CuuC* take these plural as follows:

| | CiiCaan | ʔaCCaaC | CiCaaC | CuCuuC | CiCaCat | Collective percentage |
|---|---|---|---|---|---|---|
| CaaC | 33% | 30% | 10% | 3% | 3% | 79% |
| CiiC | 4% | 62% | 8% | 0 | 12% | 86% |
| CuuC | 26% | 74% | 0 | 0 | 0 | 100% |

*Table 4-13.* Percentage of biconsonantal stems as taken by ʔaCCaaC, CiCaaC, CiiCaan, CuCuuC and CiCaCat.

Based on the percentage in Table (4-13) above, all three biconsonantal singular stems show a preference for the plural pattern *ʔaCCaaC*, but they vary in their degree of preference. For singulars of *CaaC* type, the preferred plural form is *CiiCaan*, followed *ʔaCCaaC*, which is followed by *CiCaaC*. For *CiiC* and *CuuC*, the most favored plural pattern is *ʔaCCaaC*. The second most favored plural for *CuuC* singulars is CiiCaan.

Vowel quality polarity[1] seems to play some role in the selection of plural forms by these singulars. Singulars with a low vowel (i.e. *CaaC*) tend to take plurals of *CiiCaan* and *CiCaaC*

[1] Vowel quality polarity here refers to the alternation in the selection of the singular stem by a plural pattern that tends to be expressed through polar alternation vowel quality.

types which have high vowel /i/. By the same token, singulars with high vowels (i.e. *CiiC* and *CuuC*) are overwhelmingly mapped to ʔaCCaaC plurals which have a low vowel /a/.

All the plural patterns that are taken by the biconsonantal stems have three *C* positions. Thus, biconsonantal stems have to expand in size using filler consonants to fill the empty *C* position in the output template. When consonants from the singular stem are mapped to *C* positions in a certain template, segments from the stem input will first fill the empty *C* positions at the edges of the template with the first consonant filling the *C* at the left edge and the second consonant extending to fill the *C* at the right edge. A glide (/j/ or /w/) will be used to fill the remaining *C* at the middle. The type of the glide that will be selected seems to depend on long vowel quality. Singulars with front long vowel /ii/ will use a palatal glide /j/ while singulars with back long vowel /u/ or /a/ will select a labiovelar glide /w/.

Four of the five plurals (*CiiCaan, ʔaCCaaC, CiCaaC, CuCuuC*) that are selected by the biconsonantal stems are templates that belong to the light stem group or similar to one of these templates. The distinctive feature of this group of templates is that they are overwhelmingly favored by triconsonantal stems with no long vowels. Biconsonantal stems with long medial vowel share this attribute with triconsonantal stems. Hence, based on their selection of the shape their output templates, it is justified to classify the biconsonantal and triconsonantal under the same category of light stems.

### 4.5.2 Heavy stem

**CvvCCvC, CvCCvvC, CvCCvC > CaCaaCiiC, CaCaaCiC, CaCaaCaa, CawaaCiC**

Stems with four consonants take one of the four patterns above. The stems in this group vary in their syllabic shape, but the common denominator between them is that the stem most often has four consonants. Given the difference in the number of segments compared to those in

the light and middle weight stems, this group is described as heavy stems. Some nouns have two or more plural forms from this set. Examples of each plural pattern is given in Table (4-14) below:

| | | | | |
|---|---|---|---|---|
| CuCCaaC | | | | |
| bustaan | basaatiin | | | “garden" |
| CaCCaaC | | | | |
| saʕdaan | saʕaadiin | | | “monkey” |
| sˁaħraaʔ | sˁaħaarij | | | “desert" |
| CiCCaaC | | | | |
| minqaaʃ | manaaqiiʃ | | | “chisel” |
| sirdaab | saraadiib | saraadib | | “tunnels” |
| mirʔaat | maraajaa | | | “mirror” |
| CaCCaC | | | | |
| maksab | makaasib | | | “profit” |
| CaCiCCaC | | | | |
| wasˁijjat | wasˁaajaa | | | “will” |
| raʕijjat | raʕaajaa | | | “dependent" |
| hadijjat | hadaajaa | | | “gift” |
| CaaCiC | | | | |
| ħaasid | ħawaasid | ħussaad | ħasadat | “envious” |
| kaatib | kuttaab | kawaatib | katabat | “writer” |
| ħaamil | ħawaamil | | | “pregnant” |
| ʕaaðil | ʕuððaal | ʕawaaðil | | “critic” |
| ʕaalam | ʕawaalim | | | “universe” |
| ʃaahid | ʃawaahid | | | “incident” |

*Table 4-14.* Heavy-weight stems.

Stems in this group differ from those in other groups in that they utilize templatic morphology to derive plurals. However, instead of using a template that fits all plural, I adopt the approach that heavy stems form plural by inserting the segments from stem to a combination of template-plus-projections *CaCaaCX**(C). This method was used to analyze Moroccan Arabic (Heath 1987) and consists of two parts: an output template with specified vocalism, and a hidden part whose value depends on the phonological environment. The derivation in a sense resembles evaluating an algebraic expression which consists of a constant integer whose value is known,

and a variable who is value is hidden and depends on some factors. The output template that is used to derive plural forms for heavy stems from the template-plus-projection method is not similar to the simple iambic pattern *CaCaaC* used by prosodic circumscription approach (McCarthy and Prince 1990). Unlike prosodic circumscription in which the iambic pattern only fits segments that are isolated from the first bimoraic foot, all four segments from the input stem will be mapped to their positions in the output template *CaCaaCX**(*C*). The remaining final part of the template marked as *X** will be treated as a projection from the stem input.

After mapping the segments from the stem to their positions in the output template, we derive the remaining *X** from the template. There are three types of projections that are projected on the template: /i/, /ii/ and /aa/. The selection of any of these patterns depends solely on the phonological environment (the penultimate vowel or the presence of the feminine suffix -at). The projection at the end of the plural template will be /aa/ (and the final C at the template will be dropped) if the segment input mapped to the iambic pattern are from a heavy stem that ends with feminine suffix -at, e.g. /hadijjat/ > /hadaajaa/. If the input stem has more than four segments, and the penultimate segment is a long vowel, /ii/ will be projected on the projection variable, e.g. /bustaan/ > /basaat**iin**/. If the stem has four segments and the penultimate is not a long vowel, the projection variable will be filled by a short /i/, e.g. /maksab/ > /makaasib/.

When a triconsonantal stem such as (/ħaasid/ and /dˤamiir/) is mapped to an output template with four consonant positions, it has to extend its size by adding a filling consonant. Arabic morphology uses the glides /w/ and /j/ as well as the glottal stop /ʔ/ as default filler consonants whenever it is required to extend the size of the stem input to fit into a larger output template. The type of the filler consonant and the position where it will be inserted in the output template depends on the position of the long vowel in the singular stem. A singular stem with

long vowel in the first syllable following the initial consonant will lead to the insertion of a glide /w/ in the second output position, as in /ɟaamiʕ/ > /ɟawaamiʕ/, whereas a long vowel between the second and third consonants in the singular stem will lead to the insertion of a glottal stop in the third consonantal position in the plural template, as in /dˤamiir/ > /dˤamaaʔir/.

Formulating the shape of the template is not enough to produce the output. For numerous cases, we must also specify exactly how input segments are mapped to the template positions and projected onto the accompanying projection variables. Like Moroccan Arabic (Heath 1987), we argue that mapping in MSA also does not always follow a simple left-to-right direction. We will observe numerous examples of what is called a periphery-in pattern (Heath 1987:48) where the input segments from the rightmost and leftmost edges are mapped to their peripheral output positions. After that, the medial output positions are filled by the remaining input segments. This approach may not be important for heavy stems with four segments where the number of input segments match the number of positions in the output, but it is especially important for triconsonantal stems like /ħaasid/ or /ʕaaðil/ from Table (4-14) and for singular stems with five and more consonants, not counting the feminine suffix /-at/. Examples of the latter are given in Table (4-15):

| | | | |
|---|---|---|---|
| (A) | | | |
| **ʕankab**uut | ʕanaakib | | “spider |
| **ʕandal**iib | ʕanaadil | | “nightingale” |
| **ʔuxtˤub**uutˤ | ʔaxaatˤib | ʔaxaatˤiib | “octopus” |
| (B) | | | |
| **m**an**ɟaniiq** | maɟaaniiq | | “catapult” |
| **bar**naa**maɟ** | baraamiɟ | | “program” |
| ɣ**adˤ**an**far** | ɣadˤaafir | | “strong” |

*Table 4-15.* Stems with five and more consonants.

All the singular stems above have five consonants, and when they map to an output template with four consonant positions, they have to lose one consonant. In some cases like

/ʕankabuut/ and other examples in (4-15 A), the plural loses the final consonant in the input segment. In other cases like the examples in (14-15 B), the plural form drops a consonant from a medial position in the input.

An account for the conflict between the two cases is to argue that the examples in the two cases differ due to the mapping strategies they follow. In the examples where the consonant at the stem rightmost position is dropped, segments are mapped to their output positions following a left-to-right direction where segments fill in the template positions until all available *C* positions in the output template are filled. Once all the available positions in the output are filled, whatever remains from the singular stem is dropped (as in the case with /ʕankabuut/ > /ʕanaakib/ from (4-15 A)).

On the other hand, examples where the input segments in medial position are dropped are said to follow a periphery-in mapping. In these cases, input segments from both right and left edges are mapped onto their peripheral positions in the output template. Mapping segments continues from the edges to the middle until all positions in the output template are filled, and the remaining input segments in the medial position are automatically dropped if they cannot fit in the output template (as in /barnaamaɟ/ > /baraamiɟ/ from (4-15 B)).

The phonological reason behind selecting one mapping strategy over the other is not clear. However, the singular stems that lose their final consonant during plural formation differ in their syllabic structure from the stems that lose a medial segment during plural formation. /ʕankabuut/, /ʕandaliib/, and /ʔuxtˤubuut/ have the initial four consonants not separated by long vowel, while in barnaamaɟ the third and fourth consonants are separated by a long vowel.

It can be argued that the loss of segments from a medial stem position in the plurals like, /ɣadˤanfar/, are not due to periphery-in mapping strategy but rather due to the deletion of an

extension consonant that is added during the derivation of these stems from a quadriconsonantal stem. Thus, according to this argument, the five-consonant /ɣadˤanfar/ is said to be derived from a four-consonant perfective verb /ɣadˤfar/. When the input segments are mapped to the output template to produce the plural, the segment number has to be reduced by giving priority to the four consonants that are transferred from the original stem from which the five-consonant noun is derived. It could be argued that /n/ in /ɣadˤanfar/ is not part of the quadriconsonantal stem from which /ɣadˤanfar/ is derived, and hence is dropped when five-consonant /ɣadˤanfar/ is mapped to a plural template with four *C* positions. However, this account does not lend itself to words like, /manɟaniiq/ and /barnaamaɟ/, which are not derived from quadriconsonantal stem, yet the /n/ is dropped during the plural formation.

Another piece of evidence in support for the argument of periphery-in mapping strategy is in the examples of triconsonantal stems that are inserted in a four-segment plural templates. When segments of a triconsonantal stem are inserted into a template with four *C* positions, a vacant *C* position will be created in the output template, and a filler glide has to be used to fill in the empty *C*. The position of the empty *C* will determine the mapping strategy followed. If plural formation follows a periphery-in strategy, the plural will be formed by first mapping input segments to the *C* positions at the edges and continues mapping segments from the edges to the middle until all available input segments are used. As a result of the periphery-in mapping, an empty *C* position will be created in the middle of the output template as in /ħaCaasid/ and /ʕaCaaðil/. The empty *C* position will eventually be filled by the default glide producing the correct plural form /ħawaasid/ and /ʕawaaðil/ from the triconsonantal stem /ħaasid/ and /ʕaaðil/. A left-to-right mapping, on the contrary, will fail to produce the correct plural form, as the mapping of segments in a left-to-right fashion will leave the empty *C* at the right end of the

output template, producing the ill-forms */ħasaadiC/ and */ʕaðaaliC/. Hence, the mapping strategy that Arabic follow to form plural follows a periphery-in method.

#### 4.5.2.1 The semantic group associated with CaCaaCiCat

There is a group of singular stems with four or more segments that form the plural by taking the pattern *CaCaaCiCat*. What makes this group stand on its own from the rest of heavy stems is that it includes nouns and adjectives with specific meanings. So the selection of the plural template for this special group of heavy stems is based on their semantics and meaning. Examples illustrating this are given below:

| | | | |
|---|---|---|---|
| CuCCaaC | | | |
| ʔustaað | ʔasaatiðat | ʔasaatiið | “professor” |
| mutˁraan | matˁaarinat | matˁaariin | “archbishop” |
| CaCCaaC | | | |
| ʃammaas | ʃamaamisat | | “deacon” |
| ɟabbaar | ɟabaabirat | | “titanic” |
| CiCCaaC | | | |
| ʕimlaaq | ʕamaaliqat | | “giant” |
| simsaar | samaasirat | | “broker” |
| CaCCaC | | | |
| ɟahbað | ɟahaabiðat | | “genius” |
| CiCCaC | | | |
| fitˁħal | fatˁaaħilat | fatˁaaħil | “knowledgeable” |

*Table 4-16.* Stems associated with CaCaaCiCat.

Based on the stems in Table (4-16), the template *CaCaaCiCat* is almost restricted to three semantic categories of nouns and adjectives. The first group is the nouns that refer to names of jobs and professions. The second is for adjectives that describe outstanding cerebral and physical abilities. The last group refers to the names of nationality and places of origin.

The pattern *CaCaaCiCat* share a great amount of similarity with the heavy stems templates, *CaCaaCiC* and *CaCaaCiiC*. There is also an overlap between the templates *CaCaaCiCat* and the two plural templates, *CaCaaCiC* and *CaCaaCiiC*, in terms of the stem

inputs associated with them. Most of the stems that are associated with *CaCaaCiCat* allow the formation of plural with the template *CaCaaCiC* or *CaCaaCiiC*. In the same way heavy stems select their output template, stems that are associated with CaCaaCiCat select between *CaCaaCiC* or *CaCaaCiiC* depending on the prosodic shape of the stem input. So, out of the singular stems that are associated with *CaCaaCiCat*, those that take *CaCaaCiiC* are the nouns with a penultimate long vowel (e.g. /ʔustaað/ > /ʔasaatiið/, /ʃammaas/ > /ʃamaamiis/), whereas stems that forms secondary plural by taking *CaCaaCiC* are the quadriconsonantal singular nouns with no long vowels like (e.g. /ɟahbað/ > /ɟahaabið/, /fitˁħal/ > /fataaˁħil/).

Based on the similarity in the prosodic structure between these two plural templates, and the fact that they share numerous input, it can be argued that *CaCaaCiCat* is an optional choice for a subgroup of heavy singular stems that belong to the plural patterns *CaCaaCiC* and *CaCaaCiiC*. The subgroup includes nouns and adjectives that share certain semantic properties. (The subgroup of these stems represents a semantic category that refer to names of professions and nationalities, adjectives that describe outstanding cognitive and physical traits, adjectives and nouns of nationalities).

#### 4.5.2.2 Adjectives with singular pattern CaCCaan (M) and CaCCaa (f)

The majority of stems associated with the template *CaCaaCaa* (41%) are quadri-consonantal singular nouns with a feminine ending (e.g. hadijjat). These nouns, however, seem to share the template *CaCaaCaa* with another group of singular adjectives that also form the plural by mapping the input segments to this template. The patterns of the adjectives are *CaCCaan* for masculine and *CaCCaa* for feminine. Examples of this group is given in Table (4-17):

| | | |
|---|---|---|
| sakraan(m)/sakraa(f) | sakaaraa | “drunk” |
| ʕatˁʃaan(m)/ʕatˁʃaa(f) | ʕatˁaaʃaa | “thirsty” |
| θaklaa | θakaalaa | “widowed” |
| sahraan | sahaaraa | “awake” |
| ħaznaan | ħazaanaa | “sad” |
| ɟawʕaan | ɟawaaʕaa | “hungry” |

*Table 4-17.* Adjectives with singular pattern CaCCaan (m) and CaCCaa (f).

#### 4.5.2.3 Triconsonantal stems that take heavy-stem template (CaCaa…)

In addition to the four-consonant stems discussed in 4.5.2, there is a subgroup of triconsonantal singular nouns that form plural by taking templates associated with heavy stems. Besides having three radical consonants, these singular stems share the phonological property that they should have as least one long vowel embedded in either the first syllable, the second syllable or both. Based on their semantic and phonological property, these singular nouns can be put into three groups: non-human *CaCiiC* or *CaaCiC* nouns, *CaCiiC* or *CaaCiC* stems with feminine ending and stems of the type *CvvCvvC.* Below, I provide an analysis of these groups.

The first group is stems of the patterns *CvCvvC* and *CaaCiC* that refer to non-human beings. According to the corpus data analyzed here, 79% of the 38 *CvCvvC* and 87% of 83 *CaaCiC* stems are non-human. The same conclusion was reported by Levy (1971), where the *CvCvvC* and *CaaCiC* stems that take *CaCaaCiC* share the semantic property of referring to non-human entities while the human singulars of the same patterns form plurals by taken different set of broken plural patterns. As discussed in the previous section, in order to fit into the four consonant template, a triconsonantal singular noun that is treated as a heavy stem will extend its size by adding filler glide /w/ or /j/ if the stem pattern *CaaCiC* stems and /ʔ/ if it is *CaCiiC*, and the mapping of these segments to their positions in the output template will proceed following a periphery-in pattern.

The second group of triconsonantal stems that take plural patterns associated with heavy stems is also nouns of the patterns *CvCvvC* and *CaaCiC* but with a feminine suffix ending /-at/. Like the previous group, nouns that designate non-human beings represent the majority of the *CvCvvC* and *CaaCiC* stems. (Review of the *CvCvvCat* and CaaCiCat stems that occur in the data demonstrates that non-human nouns constitute 96% of the 159 *CvCvvCat* stems and 84% of the 108 *CaaCiCat* stems). When the segments from the input stems in this group are mapped to their output positions, only the three radical consonants from the stem will be mapped to the output template, and the feminine suffix is ignored. As discussed in the Heavy stem, mapping segments to their positions in the plural template will proceed from the edges to the middle (following a periphery-in method) creating a void in the word-medial position that will be filled by one of the default filler consonants in accordance with the type of the singular stem.

The last group are triconsonantal nouns with two long vowels. This is a small group of stems, with a total of 2. As with other triconsonantal stems, the stem segments will be mapped to their positions in the output template, and filler will be added in accordance to the location and the long vowel in the singular stems. Given the important role that the long vowel in the stem plays when a triconsonantal stem is mapped to heavy template and given that the stem in this group has two long vowels, it is important to be made clear which long vowel will determine the type and the location of the filler consonant. According to the examples in the data, when a three-consonant stem has two long vowels like the nouns in this group, the priority is given the first long vowel. Hence, the stem is treated like the triconsonantal stem with long vowel in the first syllable, leading the second *C* position in the plural template to be filled by a glide.

### 4.5.3 Middle weight

There is a distinct weight class devoted to singular stems with three consonants and a long vowel. Given the shape and size of the singular inputs, I refer to this group as the middle class. Even though they form their own group, they seem to fall in the middle of the other stem groups. Stems in this group match those in the light weight class in the sense that they have three consonants, but they are close to heavy stem in the sense that they keep a size of four segments minimum (including consonants and long vowels). I divide the middle weight class into two subgroups: middle weight stems of *CvCvvC* shape, and those whose pattern is *CvvCvC*.

#### 4.5.3.1 Middle weight stems with CvCvvC shape

**CvCvvC > ʔaCCiCaaʔ, ʔaCiCCaaʔ, CuCuC, ʔaCiCCat, ʔaCCiCat, CuCaCaaʔ, CuCCaan**

Singular stems with three consonants and a penultimate long vowel form the plural by selecting one of the above patterns. Examples of this class are given in Table (4-18) below:

| | | | |
|---|---|---|---|
| CaCiiC | | | |
| sˤadiiq | ʔasˤdiqaaʔ | sˤudqaan | “friend” |
| ħabiib | ʔaħibbaaʔ | ʔaħibbat | “lover” |
| rafiiq | rufaqaaʔ | | “companion” |
| ʔamiir | ʔumaraaʔ | | “prince” |
| waziir | wuzaraaʔ | | “minister” |
| kaθiib | kuθbaan | | “sand hill” |
| ʕariis | ʕursaan | | “bridegroom” |
| xaliiɟ | xulɟaan | | “bay” |
| xaliil | xullaan | | “friend” |
| CaCuuC | | | |
| rasuul | rusul | | “messenger” |
| ðaluul | ðulul | ʔaðillat | “submissive” |
| ʕamuud | ʔaʕmidat | | “pillar” |
| CaCaaC | | | |
| ɟanaaħ | ʔaɟniħat | | “wing” |
| ʕajaar | ʔaʕjirat | | “bullet” |
| ɟabaan | ɟubanaaʔ | | “coward” |
| CiCaaC | | | |
| ɟihaaz | ʔaɟhizat | | “device” |

| | | |
|---|---|---|
| miθaal | ʔamθilat | “example” |
| wiʕaaʔ | ʔawʕijat | “container” |
| CuCaaC | | |
| duxaan | ʔadxinat | “smoke” |
| zuqaaq | ʔaziqqat | “path” |
| ʃuɟaaʕ | ʃuɟʕaan | “brave” |

*Table 4-18.* Stems with the CvCvvC shape and their plural patterns.

The syllabic structure of singular stems in Table (4-18) above are good examples of the most common input patterns that are associated with these plurals. As indicated by percentage, the distribution of the data between the most frequent singular types within each of the nine plural forms is illustrated below:

| Output | CaCiiC | CaCuuC | CaCaaC | CiCaaC | CaaCiC | CuCaaC |
|---|---|---|---|---|---|---|
| ʔaCCiCaaʔ | 24% | 0 | 0 | 0 | 0 | 0 |
| ʔaCiCCaaʔ | 94% | 0 | 0 | 0 | 0 | 0 |
| CuCuC | 24% | 12% | 5% | 30% | 2% | 0 |
| ʔaCiCCat | 52% | 9% | 0 | 26% | 0 | 9% |
| ʔaCCiCat | 12% | 3% | 23% | 41% | 0 | 8% |
| CuCaCaaʔ | 90% | 0 | 1% | 0 | 8% | 0 |
| CuCCaan | 38% | 0 | 0 | 0 | 12% | 0 |

*Table 4-19.* Percentage of the plural patterns in middle-weight group (CvCvvC) as taken by the most frequent singular stems.

As indicated by Table (4-19), the singulars with *CaCiiC* pattern are the most preferred type of input for *ʔaCCiCaaʔ, ʔaCiCCaaʔ, ʔaCiCCat*, *CuCaCaaʔ* and *CuCCaan*. Singulars with *CaCiiC* and *CiCaaC* are roughly equally favored by plurals *CuCuC*. For *ʔaCCiCat* plurals, *CiCaaC* is the most preferred type of input, *CaCaaC* is the second most favored, followed by *CaCiiC*.

The table above shows the most preferred singular patterns for each of the ten plural forms. But what are the preferred types of output for these singular stems? The preferred plural forms, as indicated by percentage, for these basic singular patterns are given in the following table:

| Input | ʔaCCiCaaʔ | ʔaCiCCaaʔ | CuCuC | ʔaCiCCat | ʔaCCiCat | CuCaCaaʔ | CuCCaan |
|---|---|---|---|---|---|---|---|
| CaCiiC | 2% | 6% | 7% | 4% | 5% | **30%** | **6%** |
| CaCuuC | 0 | 0 | **50%** | 10% | 15% | 0 | 0 |
| CaCaaC | 0 | 0 | 9% | 0 | **55%** | 2% | 0 |
| CiCaaC | 0 | 0 | 28% | 7% | **49%** | 0 | 2% |
| CuCaaC | 0 | 0 | 0 | 11% | **42%** | 0 | 5% |

*Table 4-20.* Percentage of singular stems of the shape CvCvvC as taken by the plural patterns in the middle-wight group.

Singulars with a pattern *CaCiiC* are associated with all seven plural forms, but the largest number of singulars from this type tends to go with *CuCaCaaʔ* (30%). The majority of singulars of *CaCuuC* type (50%) takes *CuCuC* as the output template. *ʔaCCiCat* is the preferred output template for singulars with *CaCaaC* (55%), *CiCaaC* (49%), and *CuCaaC* (42%).

Although singular stems in this group show a great amount of overlap in their selection of output templates, we can identify some tendencies and broad patterns in terms of how those templates are selected. The first pattern is that stems with long round vowel as in *CaCuuC* tend go with the template with round vowels *CuCuC*. Second, triconsonantal singular stems with long high vowel /ii/ in the second syllable are mapped to plurals of *CuCaCaaʔ* type while triconsonantal singular stems with a low long vowel before the third consonant show tendency to form plural with *ʔaCCiCat*.

Plural templates *ʔaCiCCaaʔ*, and *ʔaCiCCat* are used for singular stems of the pattern *CvCvvC* with identical consonants at *C2* and *C3*. They also share great similarity with the other templates *ʔaCCiCaaʔ*, and *ʔaCCiCat*. So, they seem to behave as counterparts of the non-geminated patterns *ʔaCCiCaaʔ*, and *ʔaCCiCat*, that are used for singular stems of the pattern *CvCvvC* with no identical consonants at *C2* and *C3*. Both Levy (1971) and McCarthy and Prince (1990) considered the patterns *ʔaCiCCaaʔ*, and *ʔaCiCCat* to be derived from underlyingly non-geminated patterns *ʔaCCiCaaʔ*, and *ʔaCCiCat*. So, according to these studies, the plural forms /ʔatˤbibaaʔ/ “doctor”, and /ʔaħbibat/ “lover” are surfacing as [ʔatˤibbaaʔ], and [ʔaħibbat].

Given the role that semantic qualities play in the plural formation for singular stems with specific meaning in the light and heavy weight group, I examined whether semantic properties of the singular stems influence the selection of the plural template for the nouns in the middle weight group. My investigation, as shown in Table (4-21) revealed that the designation of human entities does influence the selection of the plural pattern among *CvCvvC* stems:

| | Total | Human designating nouns | percentage |
|---|---|---|---|
| **ʔaCCiCaaʔ (ʔaCiCCaaʔ)** | 38 | 36 | 95% |
| **CuCaCaaʔ** | 96 | 95 | 99% |
| **CuCuC** | 82 | 9 | 10% |
| **ʔaCCiCat (ʔaCiCCat)** | 115 | 7 | 6% |
| **CuCCaan** | 18 | 8 | 44% |

*Table 4-21*. Percentage of the Stems that designate human and non-human in the middle-weight group (CvCvvC).

The patterns which serve as plurals for human nouns are *ʔaCCiCaaʔ* and *CuCaCaaʔ*. Singular stems that describe human entities represent 95% of the 38 singular stems that take *ʔaCCiCaaʔ* plural and 99% of the 96 stems associated with *CuCaCaaʔ*. For non-human nouns, the plural patterns are *CuCuC* and *ʔaCCiCat*, where 90% of the 82 stems linked to *CuCuC* and 94% of the 96 singular nouns associated with *ʔaCCiCat* are non-human. The pattern *CuCCaan* is evenly distributed between human (44%) and non-human nouns (46%). This is consistent with the result from previous research (Levy 1971). However, Levy prefaced the results by saying that "the conclusions should be taken as tentative in view of small number of words in any group, the large number of exceptions and the conflicting evidence sometimes presented by the dictionary and the informants" (p 44).

As described in Section (4.5.2.3), a number of *CvCvvC* stems that designate non-human beings form plural by taking *CaCaaCiC* template associated with heavy stems. When compared against the other non-human *CvCvvC* stems associated with *CuCuC* and *ʔaCCiCat*, however,

those *CvCvvC* stems that act as heavy stems remain a minority group. They make up 21% of all non-human *CvCvvC* stems. When *CvCvvC* has alternate broken plurals, these plurals are of the patterns *CuCuC* or *ʔaCCiCat*, e.g. /mudun/ & /madaaʔin/ "city"; /sufun/ & /safaaʔin/ "ship"; /sˤuħuf/ & /sˤaħaaʔif/.

None of these forms can be derived by prosodic morphology. Isolating a bimoraic foot and mapping it to an iambic template is not possible for this group of stems. Also, the mapping tendencies explained above do not account for all the plurals in the middle weight class. So, like the plural formation in the light stems, there is ample evidence suggesting that the derivation of plural forms for middle weight stems has to be lexical.

#### 4.5.3.2 Middle weight stems with CvvCvC shape

***CvvCvC > CaCaCat, CuCCaC, CuCCaaC***

The second subgroup of middle weight stems includes stems with three consonants, but the long vowel comes between *C1* and *C2*. These stems form plurals by taking one of the patterns above. Examples of this class are in the table below:

| CaaCiC | | | |
|---|---|---|---|
| qaasˤir | qusˤsˤar | | "minor" |
| saaðiɟ | suððaɟ | | "naive" |
| raakiʕ | rukkaʕ | | "kneeler" |
| saaɟid | suɟɟad | | "worshiper" |
| waariθ | waraθat | | "heir" |
| qaatil | qatalat | | "killer" |
| ħaafiðˤ | ħafaðˤat | ħaffaaðˤ | "keeper" |
| saaħir | saħarat | | "sorcerer" |
| ħaasid | ħasadat | ħassaad | "envious" |
| kaatib | kuttaab | katabat | "writer" |

*Table 4-22.* CvvCvC singular stems and their plurals.

The dominant type of singulars that form plural by taking *CaCaCat*, *CuCCaC* or *CuCCaaC* are of the type *CaaCiC*. *CaaCiC* singulars constitute 88 percent of the singulars that

take the plural pattern *CuCCaC*. For the plural pattern *CaCaCat*, *CaaCiC* singulars represent 87 percent of the singular patterns. Of course, there are some singular stems that are linked to these plurals and are not *CaaCiC*. However, these are rare and occur no more than once or two times.

I investigated the influence of semantic properties, e.g. designation of human entities, on the selection of plural template by the nouns in this group. I found that semantic properties, namely designation of the human entities, do influence the selection of the plural template by the singular nouns in this group. As explained in the previous section (4.5.2.3), a subgroup of the *CaaCiC* stem take the plural pattern *CaCaaCiC* which is normally associated with heavy stems. The distinction between *CaaCiC* nouns that act as heavy stems and the other *CaaCiC* nouns here is related to the designation of human entities. The plural patterns for human stems are *CaCaCat*, *CuCCaC* and *CuCCaaC* while non-human stems take the plural pattern *CaCaaCiC*. Human *CaaCiC* nouns make up 90% of the singular nouns that take *CaCaCat*, *CuCCaC* and *CuCCaaC*, while non-human *CaaCiC* stems constitute 87% of the nouns that take *CaCaaCiC*.

The shape of *CaaCiC* stems that are linked to the plural patterns in this group is also the form of the active participle of verb form I. Levy (1971) claims that all rational *CaaCiC* stems in her data-set are lexicalized active participles that lack the meaning transfer of true active participle. According to Levy, lexicalized active participles form plural by taking the broken plural patterns *CuCCaaC*, *CaCaCat*, *CuCCaC*, but when a given noun is a true active participle, it almost always form plural by taking the sound plural. However, there are some violations to this claim. As shown in Table (4-22), /raakiʕ/, /saaɟid/, /waariθ/, /qaatil/, /ħaafidˁ/ and /saaħir/ are true active participles yet they all take the plural templates *CuCCaaC, CaCaCat, CuCCaC*.

The *CaaCiC* stems that are linked to the plural patterns in this group belong to the morphological category of active participles. So, it is possible to regard active participles as

forming their own category of plural within the nominal plural system in Arabic, in which membership to this group is based on the semantic and morphological properties of the singular stems.

#### 4.5.3.3 Commentary on middle weight stems:

The analysis of middleweight stems has identified two types of stems: *CvCvvC(at)* and CaaCiC. Each type includes some semantically motivated subclasses. In this section, I provide a discussion of these subgroups and describe the status of middleweight stems after the semantic subgroups and residues are ruled out.

The plural formation for singular stems of the shape *CaaCiC* can be singled out as semantically motivated pluralization. Almost all of these stems, like the examples in Table (4-22), are 'lexicalized' active participles. Those with human reference are agentives, and almost exclusively take the agentive plural patterns *CuCCa(a)C* or *CaCaCat*. The remaining *CaaCiC* stems that refer to nonhuman entities form plurals by taking the regular heavy plural pattern *CaCaaCiC*. If the human agentives are excluded on grounds of having semantically restricted plurals, the remaining *CaaCiC* and other *CVVCVC* stems can be analyzed as heavy.

Singular stems of the type *CvCvvC(-at)* can also be further divided into subcategories based on their semantic characterizations. Going back to Table (4-18), *CvCvvC(-at)* stems can be split into two subcategories. There are those that are either adjectives or deadjectival nouns such as [ʔamiir] "prince" and [ħabiib] "lover" , and those that are underived nouns, such as [ɟihaaz] "device" and [rasuul] "messenger". The deadjectival nouns have their own plural patterns, namely *ʔaCCiCaaʔ* and *CuCaCaaʔ*. These plurals are also used for adjectives, as in [faqiir] "poor" and [qariib] "close" with plurals [fuqaraaʔ] and [ʔaqribaaʔ]. If deadjectival *CvCvvC(-at)* nouns are excluded on grounds of having adjectival plurals, this leaves the remaining underived

*CvCvvC(-at)* nouns that take the plural *ʔaCCiCat* and *CuCuC*, which range from human to inanimate. There is a strong case for taking underived *CvCvvC(-at)* stems as the true middleweight stems and *ʔaCCiCat* and *CuCuC* as the genuine middleweight broken plurals.

### 4.5.4 Special group of adjectives of injuries and damages

The plural template *CaCCaa* is almost exclusively used by adjectives that describe physical injuries, damages, and victimhood. In addition to the similarity in their semantic properties, these adjectives seem to have some similarity in their prosodic structure. Almost all the adjectives that take this plural pattern have three consonants and a penultimate long vowel. Examples of these adjectives are given in Table (4-23):

| | | |
|---|---|---|
| mariidˤ | mardˤaa | “sick” |
| qatiil | qatlaa | “killed” |
| sˤariiʔ | sˤarʔaa | “fallen in battler” |
| ɟariiħ | ɟarħaa | “inured” |
| ʔasiir | ʔasraa | “captive” |
| ɣariiq | ɣarqaa | “drowned” |
| majjit | mawtaa | “dead” |

*Table 4-23.* Singular adjectives of injuries.

I argue that the shape of the singular stem and semantic category they belong to both play important roles during the triage process where the stem inputs are presumably sorted into their weight class. Some of the singular stems like the adjectives in this group do not belong to a weight class but rather they form their own group based on their semantics. It can also be argued that the selection of this plural pattern depends on the two factors (shape of singular stem and semantic content) as the pattern of the singular stems of these adjective appear to be consistently of the type *CaCiiC*.

### 4.5.5 Triconsonantal with Feminine ending

***CvCCat > CvCaC***

Triconsonantal stems with a feminine ending -at but no long vowel most often form plural by taking the plural pattern *CvCaC*.

| | | |
|---|---|---|
| CuCCat | | |
| ɣurfat | ɣuraf | “room” |
| sˁudfat | sˁudaf | “coincidence” |
| buqʔat | buqaʔ | “spot” |
| CiCCat | | |
| bidʔat | bidaʔ | “innovation” |
| xirat | xiraq | “rug” |
| CaCCat | | |
| dawlat | duwal | “country” |
| ʃaqqat | ʃuqaq | “apartment” |

*Table 4-24.* Singular stems of the shape CvCCat.

As illustrated by the table above, the quality of the first vowel in the stem input is transferred to the output template. The exception to this generalization is stems with /a/, as indicated by /dawlat/ and /ʃaqqat/, which are also the least frequent among the three types of stems. Stems with *CaCCat* type take the template output *CuCaC*.

### 4.5.6 Conclusion of the analysis of stem weight in Arabic broken plural

In this section, I divide the broken plural forms into three groups based on the additive weight of their stem input. Stem weight is expressed mainly by the number of segments/slots (consonants and long vowels) in the stem input. The results of the qualitative analysis of productivity and regularity demonstrated by the broken plural forms showed that classifying Arabic plural based on the weight of the stem input is possible. Stem weight, however, is not comprehensive enough to predict the plural pattern by itself. Rather, its role can be viewed as a (quasi)well-formedness condition specified by the plural template on potential input forms.

(Weight in Arabic plural system works as triage concerned with the syllabic well-formedness of the stem inputs). A stem input that will take a certain plural template output has to meet a minimum number of segments specified by the weight condition in weight group to which a plural template belongs. Further, several stem inputs that are mapped to a common weight group, will form pockets or clusters based on the weight group to which their templates belong. The next section will test this notion of clustering of singular stems based on their weight (or weight category of their plural templates) by implementing a clustering analysis of the singular inputs in a multi-dimensional data. The question is how the pattern of clustering of the singular stems would look if a machine learning algorithm does the classification.

### 4.6 Clustering analysis by K-means and PCA

The qualitative analysis shows a possible grouping of plurals based on the weight of the stem input. This grouping is a hidden structure in the data. We have evidence from the qualitative analysis of the role of stem weight on plural patterns that this underlying pattern exists in the data. How can we validate these findings by using a computational or quantitative approach? It is possible to treat weight as a latent variable, run a clustering algorithm and let the algorithm decide how the plurals are clustered based on the similarity of their singular inputs. We then can validate the results of qualitative analysis by comparing the results of the qualitative analysis with the predictions of the clustering algorithm.

Clustering is a technique of dividing data into groups based on underlying patterns in the data. When we cluster observations, we want observations in the same group to be similar and observations in different groups to be dissimilar. So, clustering allows us to identify which observations are alike, and potentially categorize them accordingly. It is an unsupervised machine learning algorithm since it attempts to group the data without being trained on labeled

data. The clustering algorithm that we use in this analysis was the K-means clustering algorithm. The K-means algorithm groups similar data points together and discover underlying patterns by looking for a number of clusters (k), specified in advance, in a dataset.

K-means groups observations into clusters by quantifying a type of similarity relationship between them. In Arabic broken plurals the similarity relationship between data points will be measured based on a combination of the phonological features and the prosodic template of the singular stems. Phonological features are discrete units used to describe phonemes by categorizing them into groups called natural classes based on their phonological characteristics. In a featural system, a single phoneme will become a featural representation that consists of a series of binary and univalent features. Given their efficiency in describing phonemes, phonological features have been incorporated into metrics designed to quantify the similarity between morphological forms (Pierrehumbert 1993; Frisch et al. 2004). Plunkett and Nakisa (1997) used featural representations inserted into prosodic templates of Arabic singular stems to visualize the distribution of Arabic sound and broken plurals in a multidimensional space, quantify the distance between the plural types and then use the same dataset to train a multilayered network connectionist model to classify the types of Arabic plural forms. Therefore, the combination of featural representation and template of the stem will be used as parameters that K-means would use to quantify similarity between singular stems and group them into coherent clusters.

When singular stems are converted into featural representations, each phonological feature in the phonological space can be viewed as a dimension on which the data can be represented. The phonemic inventory in Arabic exploits up to 16 distinctive features to describe phonological contrasts between sounds. Thus, when singular stems are transformed to their

featural representations, the dataset becomes highly multidimensional. A large number of dimensions will affect the performance of the clustering algorithm at data segmentation since this large number of dimensions will introduce large amount of noise. Thus, this issue of high dimensionality needs to be resolved prior to any clustering analysis, in order to ensure that the analysis renders accurate results.

The statistical procedure of Principal Component Analysis (PCA) can be used to take a multivariate data set and simplify it by taking the principal components that capture the greatest variation in that data. It is a linear dimensionality reduction technique that takes data points in a given highly multidimensional space (such as the phonological space) and determines a smaller set of variables that contain the greatest variation between the data points. To determine these variables and reduce dimensionality, PCA takes the correlated dimensions (or in this case phonological features) and converts them into uncorrelated variables called principal components that are orthogonal to each other[2]. According to Ding and He (2009), using a reduction of dimensionality technique such as PCA prior to data clustering is a recommended practice since decreasing the number of features decreases the noise in the data, which thereby improves the performance of the clustering algorithm. Following the recommendation of Ding and He (2009), a PCA is first used to reduce the features in the multivariate data, then K-means is implemented on the reduced data to cluster the singular stems into plural groups.

In the next section, I explain how the data of singular nouns are prepared for this type of analysis.

---

[2] The statistical procedure that PCA use to determine the smaller set of variables that capture the greatest variance and reduce the dimensionality in multivariate data is called orthogonal transformation. It converts the observations on correlated variables into uncorrelated variables.

### 4.6.1 Data

The 4165 singular stems of broken plural nominals from the stem weight analysis are used to perform the clustering analysis. Since the clustering algorithm is an unsupervised learning algorithm that segments the data without having to be trained on a labeled dataset, the singular stems were not classified into weight groups. Only the stems that are analyzed in 4.5 are included in the clustering analysis, and those that are not included in the weight analysis are removed from the data. The number of stems that are removed from the data is 832 stems.

In order to investigate the clustering of the singular stems in the (phonological space), each singular stem has to be represented as a feature vector following the method of the template system that Plunkett and Nakisa (1997) outlined in their simulation study of Arabic pluralization. With a slight modification to Plunkett and Nakisa's method, the data of singular nouns were transformed as follows. First, the phonemes in the singular stem are aligned to a left-justified **14**-slot template, consisting of alternating consonants and vowels as CVCVCVCVCVCVCV. The slots are filled from left to right with consonants placed in C slots and vowels in V slots. Whenever the stem contains two consecutive vowels or consonants, an empty slot is inserted between them. The phonemes in the slot-based template representation are then converted into featural representations. Featural representations of all Arabic phonemes are in Table (4-25):

| | LB | LD | D | AL | PL | VR | PH | GL | N | C | V | HI | FT | SP | F | AP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| b | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 |
| m | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 |
| f | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| t | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 |
| d | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 |
| θ | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| ð | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 |
| tˤ | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 |
| dˤ | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 |
| ðˤ | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 |
| n | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 |
| r | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 |
| s | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| z | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 |
| sˤ | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| ʒ | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 |
| ʃ | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| j | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 |
| k | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| x | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| ɣ | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 |
| w | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 |
| ħ | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| ʕ | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 |
| ʔ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| h | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| a | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| i | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |
| u | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |

*Table 4-25.* Representation of all Arabic phonemes using binary features. The features are: LB = Labial, LD = Labiodental, D = Dental, AL = Alveolar, PL = Palatal, VR = Velar, PH = Pharyngeal, GL = Glottal, N = Nasal, C = Consonantal, V = Voiced, HI = High, FT = Front, SP = Stop, F = Fricative, AP = Approx.

As in Table (4-25), the feature representation in this system is demonstrated as a feature vector of **16** binary segmental features, where each feature is entered as 1 if the phoneme has it or 0 if the phoneme does not have it. Empty slots that occur between consecutive consonant or vowels or at the end of the stem are represented as a feature vector of 16 zeros. The outcome of transforming the 14 segments in the template into 16 features is that each singular stem is converted into a 224 elements vector, or (1x224) matrix. A visual demonstration of the design of the data is given in figure (4-1).

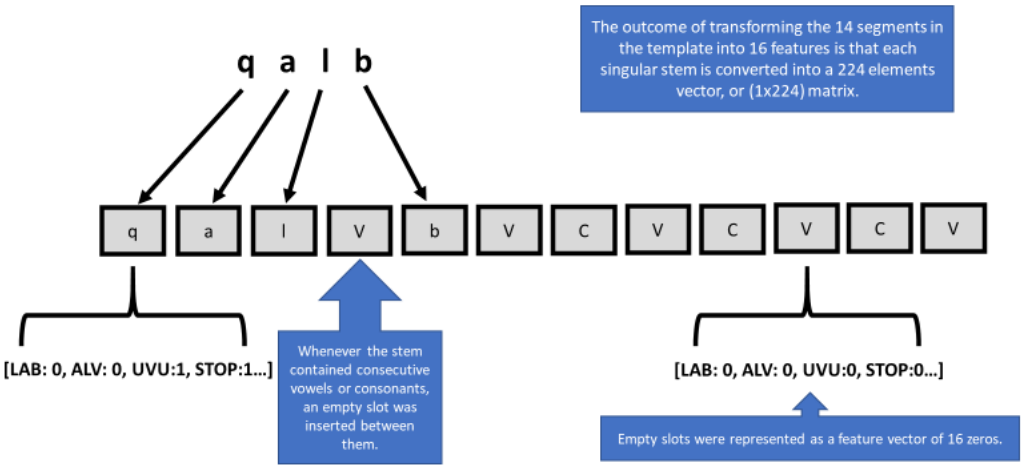


*Figure 4-1.* An illustration of the process of converting a singular stem to a feature vector.

The transformation of the data by using Plunkett and Nakisa's (1997) template system and representing the singular nouns in a fixed 14-slot template is necessary for the implementation of PCA and K-means algorithms. The algorithms require all the singular nouns to be represented as an array of numbers or a matrix of the same size. Inserting the phonemes from the singular stems into a fixed 14-slot template ensures that all singular stems would be represented as vectors of fixed length, and hence allowing for the investigation of their clustering pattern to be performed.

### 4.6.2 Analysis

The analysis is executed in two parts. In the first part, I perform PCA to reduce the number features in the dataset. As I explain in 4.6, the purpose of implementing PCA is to extract a set of variables that captures the greatest amount of variance from a multivariate data. How do we decide whether a principal component is statistically important enough to be extracted from the data? There are several criteria to make this decision. One criterion advocated

by Cattel (1966) is to obtain the Eigenvalue of each component, plot a graph of each Eigenvalue

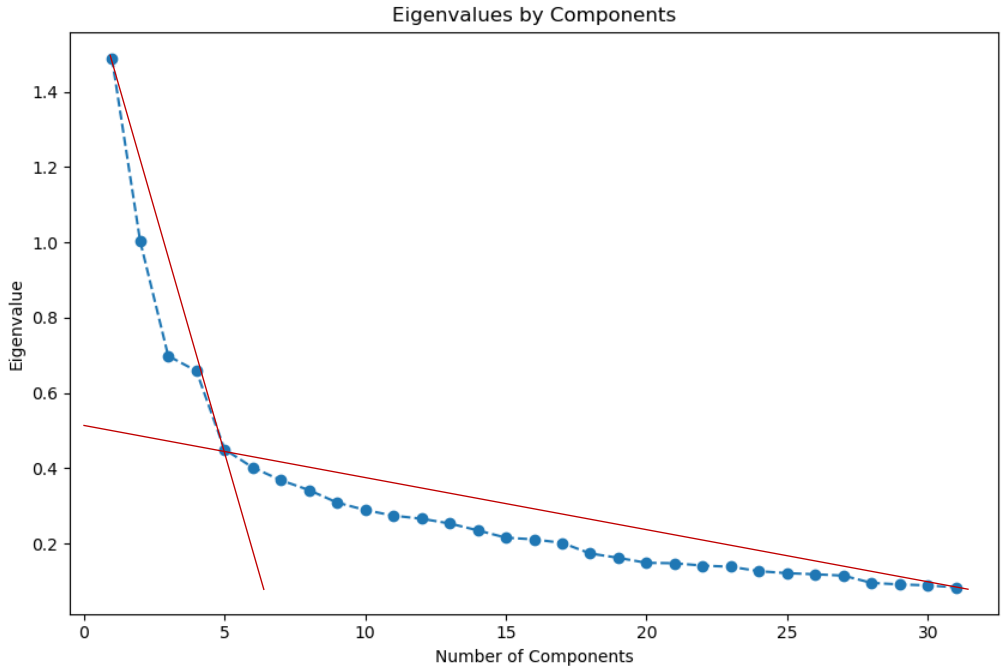


*Figure 4-2.* Plot of the first 31 components with the highest Eigenvalues on the x-axis and their Eigenvalues on the y-axis.

(y-axis) against the component with which it is associated (x-axis). Eigenvalue is a measure of the relative importance of each component, so Cattel (1966) argues that the cut-off point for selecting components should be at the point of inflexion where the slope of the line changes dramatically. Figure (4-2) shows a plot of the first 31 components with the highest Eigenvalues on the x-axis and their Eigenvalues on the y-axis. In this Figure the point inflexion occurs at the fifth point (component), therefore, the number of components that should be extracted is four.

Another method of component selection is Kaiser's criterion. Kaiser (1960) recommended retaining all factors with eigenvalues greater than 1. The idea behind this method is that eigenvalues represent the amount of variance explained by a component, and according to Kaiser (1960), 1 is a substantial amount of variance. As shown in Figure (4-2), there is only one component with an eigenvalue greater than 1, where the absolute Eigenvalues of that component

is 1.49. Thus, based on Kaiser's criterion the number of components that should be extracted is 1. Jolliffe (1986) argues that the Kaiser's criterion is too strict and may exclude important components in the data. As an alternative option to Kaiser's criterion, Jolliffe suggests retaining components with eigenvalues greater than 0.7. Based on this method the number of components that should be extracted becomes 2 as the first two components have eigenvalues of 1.49 and 1.0. The number of extracted components based on Kaiser's criterion or that of Jolliffe is lower than the number prescribed by Cattel's suggestion to extract components at the cut-off point after the point of inflexion in the Scree Plot. Following the Kaiser's Criterion and extracting the first two components explains 22% of the cumulative variance in the data, whereas including the first four components increases the cumulative variance explained to 35%. Therefore, the number of components that will be used in the K-means analysis will be restricted to the first 4 components.

Having determined the number of principal components that will be extracted, we proceed to perform the K-means clustering on the reduced data. However, we had to specify the number of clusters that the K-means algorithm should look for, the optimal value of K. The method that is used to determine the optimal value of K is the elbow method. The idea behind this method is to implement the K-means algorithm several times starting without any cluster (K = 1) and increasing the number of K by 1 at each implementation. After the implementations are completed, the Within Cluster Sum of Squares (WCSS) will be plotted against the number of clusters at each implementation. WCSS is a measure of the variability of observations within each cluster. K-means clustering partitions data into clusters such that the total within-cluster variation is minimized. Clusters should be added until adding clusters does not improve WCSS. The cut-off point at which the WCSS does not improve (or becomes a flat line) should be the

optimal value of K. A graph of WCSS against the number of clusters from 35 implementations of K-means algorithm on the data is given in Figure (4-3):

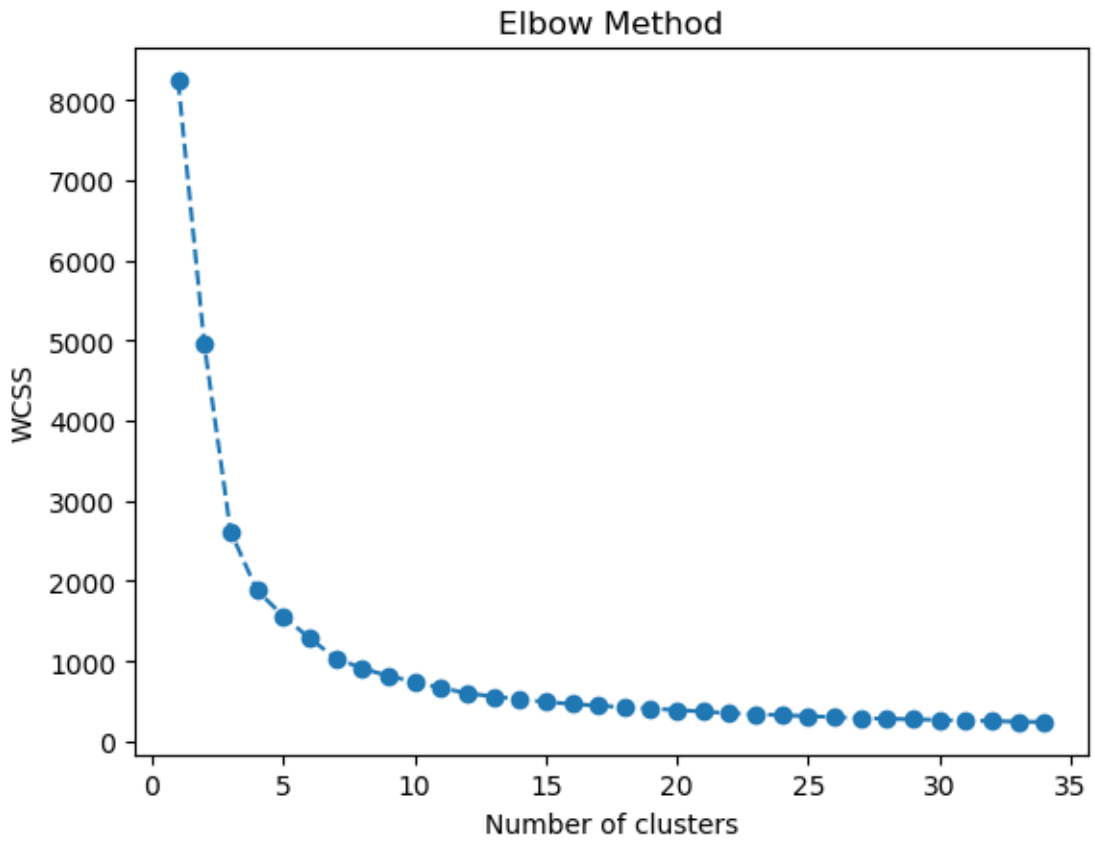


*Figure 4-3.* A graph of WCSS against the number of clusters.

According to Figure (4-3), the variance starts to become low (flat line) at 4 clusters. The cut-off point at which the WCSS stops improving is at cluster 4. So, the number of clusters based on the optimal value of K that the K-means algorithm should look for is 3.

The PCA scores of the first four principal components from PCA were entered into a K-means algorithm with an optimal K-value of 3. (That is, the number of clusters the K-means should look for is 3). Since a dimensionality reduction technique was already applied to reduce the data to the most important variables, it is possible to use the reduced space to visualize the

clustering by K-means by projecting the data onto the lower dimensional space. Figure (4-4) shows the three clusters that the data points from Arabic singular stems exhibit in a plane defined by the first two components. K-means segments the data into 3 segments that are randomly named 0, 1 and 2. The figure showed that the Arabic singular stems falls into three coherent clusters.

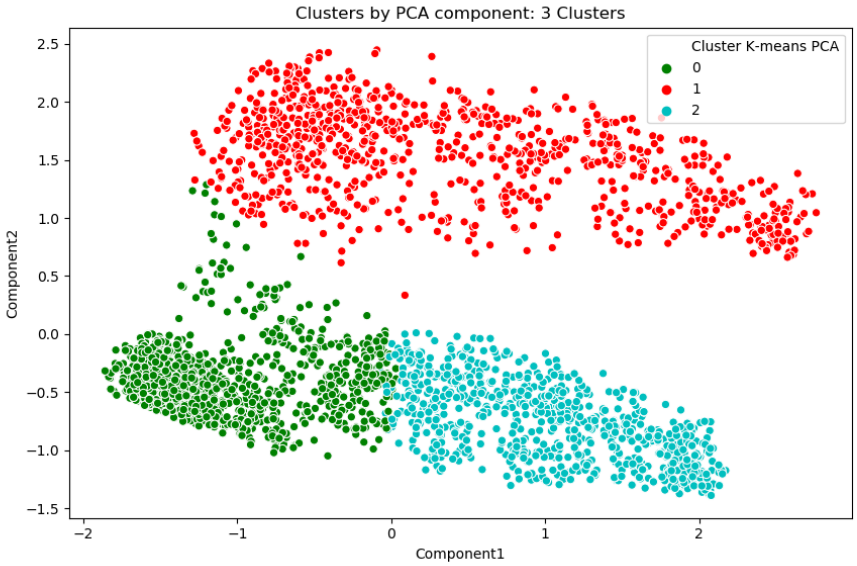


*Figure 4-4.* The segmentation of Arabic singular stems by the K-means.

To identify the underlying patterns in the data, the labels that indicate the cluster or the group assigned for each singular stem by K-means are added to a data frame that contains the plural patterns that correspond to these singular stems. So, every plural pattern could be mapped to the cluster that its singular stem is assigned to. Then, the number of occurrences of plural patterns in the three clusters is counted. Table (4-26) shows the frequency of the plural patterns in each of the three clusters.

| Template | Cluster 1 | Cluster 2 | Cluster 3 |
|---|---|---|---|
| ʔaCCaaC | 576 | 2 | 30 |

| | | | |
|---|---|---|---|
| CuCuuC | 337 | 0 | 11 |
| CiCaaC | 145 | 5 | 79 |
| ʔaCCuC | 38 | 0 | 17 |
| ʔaCCiCat | 9 | 1 | 125 |
| ʔaCCiCaaʔ | 0 | 0 | 21 |
| ʔaCiCCat | 0 | 1 | 28 |
| ʔaCiCCaaʔ | 0 | 1 | 18 |
| CaCaCat | 0 | 0 | 58 |
| CuCCaC | 3 | 0 | 30 |
| CuCCaaC | 18 | 0 | 142 |
| CuCaCaaC | 0 | 0 | 100 |
| CuCuC | 12 | 12 | 79 |
| CaCaaCaa | 10 | 58 | 5 |
| CaCaaCiC | 211 | 454 | 282 |
| CaCaaCiCat | 7 | 27 | 7 |
| CaCaaCiiC | 14 | 357 | 3 |

*Table 4-26.* The distribution of the plural patterns in the three clusters

The results are consistent with the findings from the qualitative analysis of stem weight. According to Table (4-26), patterns with the highest occurrences in cluster 1 are *ʔaCCaaC*, *ʔaCCuC*, *CuCuuC* and *CiCaaC*. These patterns are the same patterns that describe light weight in the qualitative analysis. The patterns occurring most frequently in cluster 2 are *CaCaaCaa*, *CaCaaCiC*, *CaCaaCiCat* and *CaCaaCiiC* and they are the same patterns that are associated with heavy weight stems. The most frequent pattens in the last cluster are *ʔaCCiCaaʔ, ʔaCiGGaaʔ, CuCuC, ʔaCiGGat, ʔaCCiCat, CuCaCaaʔ, CuCCaan, CaCaCat, CuCCaC, CuCCaaC*. These patterns are the plural templates associate with middle stems. Based on the frequency distribution of plural patterns in the three clusters, the underlying pattern that the K-means identifies

conforms to the notion of additive weight of the stem input that divides plural patterns into three groups based on the additive weight of their singular input.

### 4.6.3 Discussion

K-means combined with PCA is implemented on Arabic singular stems to determine the underlying clustering patterns of their broken plural templates. The results show that singular stems fall into three clusters (Figure 4-4). The analysis of the plural templates associated with each cluster (Table 2-26) shows the clustering pattern is consistent with the results from the qualitative analysis where the plural templates are classified based on the additive weight of their stem inputs.

Although the cluster analysis successfully divides the data into three clear segments, there is some overlap between these clusters. Cluster 3 which includes plural patterns linked with middle weight stems has the lowest degree overlap, since only 4 out of the 9 patterns overlap with another cluster. When these patterns are not classified in cluster 3, they tend to occur in cluster 1 that is dominated by patterns associated with light stems. Patterns in cluster 2, which is primarily associated with heavy stems, overlap roughly in equal amount with clusters 1 and 3, and the greatest overlap is attributed to CaCaaCiC. Patterns in cluster 1 almost exclusively overlap with cluster 2.

## 4.7 Summary and conclusion

The results from qualitative and computational analyses provide evidence for the role of simple additive weight on the mapping of singular input stems to plural outputs in Arabic broken plural. Plural patterns can be effectively grouped into classes based on the number of segments in their singular stems. The use of this approach of simple additive weight as opposed to the metrical approaches can also be seen in many unrelated languages. The effect of the additive on

weight on morphological processes, for example, can be seen in the verbal morphology in Tamashek (Heath 2005) and the tonal patterns on verbs in Penange (Heath 2018). The fact that these languages are unrelated provides further support for a model of morphophonological weight that adopts on additive weight of the form rather than metrical units.

The findings about the role of additive weight of the stem input on plural derivation currently lack supporting evidence from behavior studies. The clustering analysis performed on a corpus of Arabic singular-plural pairs verifies the conclusions from the additive weight model of broken plural derivation in Arabic. Yet, as far as Arabic native speakers are concerned, it is not clear how the additive weight of the singulars will influence native Arabic speakers' processing of nominal plurals. An opportunity for future research is to conduct a judgment experiment to determine whether stem weight affects speakers' decisions in the same way predicted by the current analysis.

The current assessment of the influence of stem weight on Arabic broken plurals only covers the most frequent plural patterns. Apart from the plural patterns included in the current study, there are multiple plural patterns that are excluded from the current study because they are infrequent or anomalous. These infrequent patterns and anomalies are beyond the scope of this dissertation and warrant further investigation.

The results show that it is possible to classify broken plural patterns into categories on the basis of the weight of their singular stems. One possible interpretation for the results is that stem weight functions as a (quasi) well-formedness condition imposed by the plural template on potential stem inputs. It can be viewed as a rudimentary sorting mechanism used in a complex system such as the broken pluralization in Arabic to reduce the number of inputs by eliminating the forms that violate the required weight. As an initial sorting step, the weight approach does

not map the singular stem to its exact plural form but rather the singular forms are mapped to a group that includes variants of templates. Once a singular stem is assigned to a weight group, the plural template will be determined by a number of factors that include syllabic shape of the stem (Levy 1971; McCarthy & Prince 1990; Ratcliff 1998), vowels (Ratcliff 1998), rationality or human designation (Levy 1971), and the importance of these factors varies. These factors and their contribution to determining the plural pattern for a particular singular stem is investigated in the next chapter.

# Chapter 5 Computational Analysis of the Factors Involved in Deriving Generalizations in Arabic Nominal Plurals

## 5.1 Introduction

The noun plural system in MSA poses a major challenge to morphological learnability theories for numerous reasons. First, there are large number of possible plural patterns/templates for the singular stems to select when forming plural. Wright (1988) and McCarthy and Prince (1990) listed as many as 31 productive plural patterns, but Plunkett & Nakisa (1997) argue that when the infrequent plural patterns are included, the number probably is greater than 70. Second, the process of broken pluralization in Arabic by associating singular stems to plural forms manifests a many-to-many mapping. Singulars that have the same shape do not necessarily take the same plural patterns. For examples, in [ʔamiir] > [ʔumaraaʔ] 'prince', and [sˤadiiq] > [ʔasˤdiqaaʔ] 'friend', the nouns have the same *CvCvvC* shape for the singular, yet they all take different plural patterns. Conversely, words that have the same plural patterns do not necessarily have the same shape in the singular. In [sˤifr] > [ʔasˤfaar]; [qalam] > [ʔaqlaam], for example, the nouns take the plural pattern *ʔaCCaaC,* but they all have different shapes in the singular. There seem to be no simple correlation between the singular stem and its plural pattern. Thus, the challenge is how to explain how the learning of plural formation is achieved when the plural system involves such a seemingly random and chaotic process with a many-to-many mapping between singulars and plurals. The last reason for considering Arabic broken plural as an interesting challenge for theories of morphological learning is the less understood mapping rules of singular to plural. The lack of the simple correlation between the singular and the plural

pattern makes it difficult for any attempt to lay out the theoretical basis underlying the process of deriving plural forms for singular nouns. For these aforementioned reasons, the Arabic noun plural system provides an excellent opportunity for examining key issues in the theory of morphological learnability.

Although the way in which these singulars select their plural template is far from being simple, as the mapping in some cases involve what can be described as a many-to-many relationship, previous research has pointed out major factors that capture some tendencies implied through the mapping of singular stems to their plural patterns. Among these factors, the prosodic shape of the singular stem (a.k.a. singular CV pattern) has been shown to be the primary determinant of its plural form. The CV template of singular stems is immensely effective in separating the nouns that forms plural by taking sound plural suffix from those that form plural by taking one of the broken plural templates. Both Levy (1971) and McCarthy and Prince (1990) showed that singular nouns that take broken plural templates are well defined by the shape of their CV template, whereas the sound plural is systematically observed only with a small set of word types, which include: proper nouns, transparently derived nouns or adjectives, unassimilated loans, and the names of the letters of the alphabet, (which means that plurals for these singulars are extremely predictable). Even when the singular nouns is associated with more than one plural patterns, it is possible to predict these plural patterns as the singular nouns do not take these patterns with the same preference. As shown the previous chapter, almost in all cases when the singular is linked with multiple plural forms, one of these patterns will occur frequently and, hence, become the default while the other will be used for rare cases and with low frequency, which means the plural patterns are predictable even for cases with multiple plural forms.

Vowel series or vocalism of the singular stem is another principal factor involved in pluralization of many singular nouns. Both vowel length and quality of the singular stem were proven to play major a role in in limiting number of choices of plurals that the singular stem will pick (Levy, 1971; McCarthy & Prince 1990; Ratcliffe 1998). For example, both [taqliid] "tradition" and [masɟid] "mosque" are heavy stems and have the same vowel quality but differ with their vowel length, and this difference in their vowel length influences the plural patterns they take. So, when they form plural, [taqliid] takes the plural template *CaCaaCiiC* producing [taqaliid] while [masɟid] takes the plural template *CaCaaCiC* as in [masaaɟid]. Levy (1971) as well as Ratcliffe (1998) indicated that vowel quality also affects the singular-to-plural mapping process, especially when singular nouns have the same prosodic shape but take different plural patterns. Monosyllabic masculine singular nouns is a perfect example of this case. All nouns in this subgroup of singular stems have the CV template *CVCC*, and they can take any of these plural templates: *ʔaCCaaC*, *CuCuuC*, *ʔaCCuC*, and *CiCaaC*. Predicting which plural pattern the singular noun will take on the basis of their templatic shape only seems to be impossible. However, the quality of vowel in these nouns can give a hint of the preferred plural pattern for each singular stem. Singular nouns with low vowel *CaCC* predominantly take *CiCaaC* and *CuCuuC* as their plural template, while stems with high back vowel take *ʔaCCaaC*.

Many times, the presence of geminates or weak consonants acts as indicator for exceptions that do not take the major plural pattern taken by the majority of singular stems of the same shape. As Levy (1971) and Ratcliffe (1998) pointed out, most of the time the significance of geminate and weak consonants in the singular stem is that, within a group of singular stem of the same shape, geminates and weak consonants help to highlight the exceptions that do not take the default plural pattern. In the group of light stems, for instance, the majority of singular nouns

with a *CaCC* pattern forms plural by taking the plural template *CiCaaC* whereas light stems with high back vowel *CuCC* are more like to take the plural template *ʔaCCaaC*. However, when the singular with *CaCC* shape ends in a glide in the final *C* position, they go against majority of singular stems with the same shape and take the plural *ʔaCCaaC* as in [bahw] > [ʔabhaaʔ] "lobby", making the final glide a reliable factor to predict the subgroup of *CaCC* stem that will go against the majority of *CaCC* stems and take different plural pattern.

Finally, certain semantic qualities of the singular stem have also shown that there is some "regularity" in the system beyond that defined by the CV template or the vocalism. For nouns that take sound plurals, it is well understood that both animacy and gender of the word in tandem mediate the choice between the two sound plurals (Levy 1971; Ratcliffe 1998). The sound [-uun] plural attaches to human masculine nouns with few exceptions, while the sound [-aat] plural attaches to human feminine nouns as well as non-human nouns. For nouns that take broken plurals, the analysis of broken plural in the previous chapter has revealed that when a singular stem is associated with multiple plural patterns, animacy and to a lesser extent abstractness conveyed by the word become one of the factors that determine the choice between two plural templates (Levy 1971). Animacy of the word partially determine the plural template the singular stem will select. As shown in the previous chapter, singular nouns that have the CV shape *CaaCiC* take the plural patterns *CuCCaaC* and *CaCaaCiC*, and the selection between the two is based on whether the noun designates human or non-human entities; those that designates human entities take *CuCCaaC* while all non-human designating nouns virtually take *CaCaaCiC*. Accordingly, semantic features can be a significant predictor of the plural pattern for a given noun in this group.

The study in this chapter aims at providing a computational analysis of the mapping of singular stems to plural forms to explore what types of information are relevant in making generalizations in Arabic plural system. The contribution of a series of factors, such as CV template, vowel series and semantic properties (represented in the semantic distinction between human-designating vs. non-human-designating nouns), which were reported in the literature to play a role in the singular-to-plural mapping models, on the prediction of plural patterns for singular stems will be investigated by implementing computational predictive models. To be specific, I want to build multiple K Nearest Neighbor (KNN) classifiers that use these factors to select plural patterns, then compare their performance and accuracy. This will give us an idea of the contribution of each factor on learnability of noun plural system in Arabic. First, I will give a review of previous studies that have done computational analyses of plural selection in Arabic, describe their limitations and explain how the current study will address these limitations. Next, I discuss the results of the KNN models used to predict the plural patterns based on information from the singular stem. Finally, a general discussion describes implications of the findings and possible opportunities for future research.

## 5.2 Previous Research

Three previous studies have examined predictive computational models of the Arabic noun plural system. Plunkett and Nakisa (1997) employed a multi-layered connectionist network to classify Arabic singular forms according to their plural type, as one of the 12 broken plural types or one of the 2 sound plural types, to examine the capacity of the network classifier to generalize, after training, to novel plural forms. The data-set consisted of 859 singular–plural pairs from the Hans Wehr (1976) dictionary. The broken plural accounted for 76% of the data-set, with broken patterns ranging from 1.5% to 17.5% of the data-set. The singular forms were

represented as vectors containing 16 segmental features per phoneme, which were then mapped onto a left-justified VCVCVCVCVCVCV template. Unused slots were represented as an empty vector, for a total vector of 208 features per word. The network was trained on the singular forms, using the 208-feature vectors as input units and plural type as output units.

To investigate the capacity of the multi-layered connectionist network classifier to classify the plural type of Arabic nouns on the basis of the segmental features of the singular stems, Plunkett and Nakisa performed a series of five mapping simulations, with each simulation starting at a different random seed. After 1000 epochs of training on the entire data-set, the five networks were able to learn 93.3% of trained forms. The profile of errors produced initially by the network over the 1000 epochs of training showed that the majority of errors involved overgeneralization to the most frequent plural patterns in the data. Most of the errors were broken plurals that are regularized to feminine sound plural, or sound plurals that are treated as broken plural patterns *CaCaaCiC* and *ʔaCCaaC*. As training proceeds, however, the proportion of errors caused by regularizing to sound plural (i.e. incorrect use of sound plural) decreases whereas the proportion of errors attributed to irregularization to broken plural (i.e. incorrect use of broken plural) decreases. Plunkett and Nakisa, based on the analysis of the profile of errors produced by the network classifier, made predictions about the acquisition of plural by Arabic speaking children. During the early stages of language acquisition, they predict that Arabic speaking children will most likely overregularize many of the broken plurals to the female sound plural and irregularize broken and sound plurals to broken plural patterns *CaCaaCiC* and *ʔaCCaaC*. Later in development, however, broken plural over-generalizations will constitute most of the errors.

To examine the capacity of the network classifier to generalize to novel plural forms, Plunkett and Nakisa then trained the network on 90% of the data-set and using the remaining 10% as a held-out test set. A 10-fold cross validation was implemented to ensure the results are robust to variation. This was performed by iterating the testing procedure 10 times, each time with a randomly selected 10% of the data-set used as the test group and the remaining 90% used as the training group. Then, accuracy of the model at predicting the plural patterns is averaged across the 10 iterations. For unseen forms, the network correctly classified 63.8% of forms. Importantly, the network trained on the full data-set was able to learn both trochaic and iambic broken plural patterns with good accuracy, which suggests that both of these prosodic patterns should be productive. The much lower accuracy for unseen forms suggests that the plural system is learnable, but that generalization to new forms may be a more difficult task.

Nakisa et al. (2001) employed the connectionist network from Plunkett and Nakisa (1997) in addition to a k-nearest-neighbors model and the generalized context model (GCM; Nosofsky, 1990) to evaluate the performance of these models using single-route approach in generalizing to novel forms with the performance of models that use dual-route approaches in generalizing to the same novel forms. The single-route models used only the classifier in question, while the dual-route models also had classifiers to classify broken plurals and a rule-based module that was triggered to classify input forms as sound plurals whenever the first module failed to reach a threshold of similarity to the test form (if the input forms were not recognized as broken plurals by the first module or the classifier(the associative memory component)). The data- set was much larger than in the previous study (n = 4771 singular–plural pairs) but contained only 11 broken plural patterns. The broken plural constituted 73.6% of the data-set. The single-route models achieved higher accuracy across the board on unseen forms

than the dual-route models, and the single-route GCM and connectionist network were equally more accurate than the single-route k-nearest-neighbors.

Comparing the results of the single-route classifiers and the dual-route models, Nakisa et al. (2001) found that the performance on broken plurals is essentially identical for the single- and dual-route models, as these are both handled by the classifier component (i.e. neural network) in the model. However, when they compared performance of these models on sound plurals, they ironically found that the dual-route model performs particularly poorly when generalizing to the sound plural, which is the arguably default in the Arabic plural system. This, as Nakisa et al. explains, is because membership to the sound plural is simply predictable from the phonological shape of the singular nouns and singulars which take the sound plurals cluster together in phonological space as coherently as many of the broken plural classes. Single-route models, which only use classifiers to generalize to unseen forms, thus, are able to take advantage of this characteristic (i.e. clustering of singular stems based on the similarity in the structure) in generalizing from known forms to novel forms in training. They finally concluded that generalization to unseen or novel forms in Arabic is more accurately predicted by their similarity to existing forms in the language, rather than by the operation of a default rule.

Although the early predictive models managed to predict plural patterns for singular nouns with fairly good accuracy, these early studies had empirical problems that need to be addressed. Dawdy-Hesterbeg and Pierrehumbert (2014) list three of these major limitations in these early studies. First, the data-set in both studies came from a dictionary, which comes with the disadvantage of not being representative of language in actual use, hence leading the authors to make assumptions that are not necessarily true. For example, the distribution of sound and broken plurals in the data-set in both studies represented sound plurals as a minority group, a

claim that was found to reflect unrealistic distributions of sound and broken plurals by Dawdy-Hesterberg and Pierrehumbert. Dawdy-Hesterberg and Pierrehumbert's analysis of 6597 singular-plural pairs from the Corpus of Contemporary Arabic (Al-Sulaiti 2009) found that sound plurals are statistically dominant (74% of the plural type)). A second empirical issue Dawdy-Hesterberg and Pierrehumbert found in both studies is that both studies addressed only a subset of broken plural types. Wright (1988) cites 31 common broken plural patterns, but Plunkett and Nakisa (1997) examined only 12 broken plural patterns, while Nakisa et al. (2001) examined 11 patterns. Dawdy-Hesterberg and Pierrehumbert pointed out that capturing all generalizations in the Arabic nominal plural system is unachievable with only a small subset of plural patterns and a large set of the plural patterns that occur in the system being neglected. The third limitation that Dawdy-Hesterberg and Pierrehumbert identified in these studies is the little insight into the factors that govern pluralization in Arabic. Dawdy-Hesterberg and Pierrehumbert indicated that an analysis of generalization and learnability of Arabic plural system should address these limitations.

To address the limitations in Plunkett and Nakisa (1997) and Nakisa et al. (2001), Dawdy-Hesterberg and Pierrehumbert (2014) employed the GCM as implemented in Nakisa et al. (2001) but sought to include three new contributions. First, they compared five variations of the GCM to determine the importance the contribution of three specific factors in plural selection: 1) segmental features of the singular stem, 2) gang size represented by the count of a singular-plural pair that share the same CV templates of both the singular plural forms, and 3) CV template of the singular stem. Second, they used a data-set that encompasses 37 broken plural types, covering a large range of plural patterns that were not included in Plunkett and Nakisa (1997) and Nakisa et al. (2001). Finally, they used a data-set that were collected from a

corpus and represents a more realistic distribution of broken to sound plurals, with 28.8% broken plurals in the whole data.

The full data-set that Dawdy-Hesterberg and Pierrehumbert (2014) used in their simulations consisted of 1945 pairs of singular and plural forms collected from the Corpus of Contemporary Arabic (CCA; Al-Sulaiti 2009), which contains approximately 840,000 words from different written genres. Pairs of singular and plural forms were categorized into groups called 'gangs' based on their prosodic shape, i.e. CV pattern, of the singular and plural forms of the word, constituting a total of 108 gangs. The breakdown of the 108 gangs over the three plural types is as follows: 55 singular-plural gangs taking the sound feminine plural, 16 taking the sound masculine plural and 37 taking the broken plurals. (There was one problem with the data-set they used). The words from CCA are written in Standard Arabic orthography with no diacritics or symbols that show short vowels, geminates, and semivowels [w] and [j]. As result of the lack of the diacritics, all the nouns in their singular and plural forms used in the analysis were unfortunately presented and analyzed without the short vowels, semivowels or geminates, a serious empirical limitation in the study as we will see. Dawdy-Hesterberg and Pierrehumbert recognized that the omission of these features is a limitation that needs to be addressed in future research.

Although Dawdy-Hesterberg and Pierrehumbert (2014) helped to fill an important gap in the study of generalization in Arabic plural system, their research has some limitations and empirical problems that need to be addressed. The first limitation involves the set of factors that they analyzed in determining the plural patterns for the singular. As discussed above, the process of predicting a plural pattern for a singular stem is influenced by several factors that includes the prosodic shape of the singular stem, the quality and length of stem vowels (Levy 1971; Ratcliffe

1998; McCarthy and Prince 1990), the presence of weak consonants and geminates (Levy 1971; Ratcliffe 1998), semantic qualities of the noun (Levy 1971), the size of the singular-plural group to which a singular belongs and the segmental features within the singular stem (Dawdy-Hesterberg and Pierrehumbert 2014). However, Dawdy-Hesterberg and Pierrehumbert examined the influence of only three factors on the selection of the plural. The importance of several factors involved in the selection of plural templates remains unknown. In addition to examining the well understood influence of CV template, I wish to know the extent to which vocalism and semantic qualities of the singular stem influence morphological generalization in Arabic plural system. Theoretical and traditional analysis showed that All these factors play role in the selection of plural template for singular stems and broadly in the formation of the morphological generalization in Arabic plural system, but the contribution and importance remain unknown.

A second empirical issue in Dawdy-Hesterberg and Pierrehumbert (2014) is that they used data-set that show words in undiacritized or unpointed representation, (which do not show the short vowel, semivowel and the geminates) since the data-set collected from a corpus of Arabic text written in the Standard Arabic Orthography. Unpointed orthography presents numerous issues for text analysis (see Buckwalter 1997). The so-called Standard orthography for so long has dropped the diacritics or points necessary for marking short vowels, geminate consonants and certain semivowels since the use of diacritics becomes redundant at some point when native speakers become able to predict them from the context. However, without the ability to understand the context, the words become ambiguous and the use of diacritics become necessary. A word that is written in undiacritized orthography, like in the case of the data-set used in Dawdy-Hesterberg and Pierrehumbert, hides the short vowels, geminates and semivowels, and hence any attempt to create a phonological transcription of the words based on

their undiacritized orthography will yield an inaccurate transcription. Without short vowels, for example, the two words [ʕilm] -> [ʕuluum] “science” and [ʕalam] -> [ʔaʕlaam] “flag” will be incorrectly merged into single ill-formed word /ʕlm/, that single word will presumably have two plural forms, when, in fact, this is not case at all. The same thing applies to nouns with semivowels in certain environments or geminates which will be falsely transcribed as similar to nouns without geminate or semivowels as a result of this empirical limitation. Any future analysis of the morphological generalization in Arabic plural system must use a data-set that include all the phonological information in the word.

The third limitation in Dawdy-Hesterberg and Pierrehumbert (2014) involves the size of the data-set used in their analysis. Dawdy-Hesterberg and Pierrehumbert criticized Plunkett and Nakisa (1997) for conducting their analysis of morphological generalization and learnability in Arabic nominal plurals on a fairly small-sized data-set. (They considered that to be one of the empirical issues in Plunkett and Nakisa’s study). To address this limitation in previous research, Dawdy-Hesterberg and Pierrehumbert conducted a similar analysis of Arabic plural selection on the basis of features from the singular stems on a larger data-set (n = 1384). Ironically, when the data-set from Plunkett and Nakisa (1997) and the data-set from Dawdy-Hesterberg and Pierrehumbert (2014) are divided by sound and broken plural forms, the number of broken plural forms (n = 561) Dawdy-Hesterberg and Pierrehumbert analyzed becomes smaller than the number of broken plurals (n ≈ 635) that Plunkett and Nakisa (1997) analyzed. Sample size of 500 plurals may exclude some instances of Arabic nouns that are used by Arabic speaker and are not included in the corpus. The problem of a small number of broken plural may not be representative of the realistic distribution of broken plurals, with respect of the broken plural patterns and their frequency. Also, it cannot capture all the anomalies and exceptions that occur

in the system. Using a larger size sample would be more representative and more likely to cover broader examples that occur in the system.

In this study, we employ a KNN classifier to predict plural patterns for singular noun stems to analyze morphological generalization and learnability in Arabic plural system. More importantly, the study will address the shortcomings and empirical issues from previous research by adopting the following new measures: 1) we compares the performance of multiple KNN classifiers at predicting the output to determine the effect of a range of factors, especially those ignored and understudied by previous research, including vowel length and quality, semantic qualities represented in human designation on plural selection; 2) the data-set used in the analysis is collected from a corpus that is diacritized and pointed to make sure that all words are accurately transcribed with short vowels, semivowels, and geminate; and 3) to overcome the problem of under-representing the broken plural forms, the data-set has 4098 pairs of singular and plural forms.

## 5.3 Method

### 5.3.1 Data

The same data-set used in the analysis of statistical distribution of sound and broken plurals (Chapter 2) is used to perform the computational analysis. The data-set consists of (8022) singular-plural pairs collected from a subset of Arabic Gigaword (Parker, Graff, Chen, Kong, & Maeda, 2011), which approximately contains 300,000,000 words. The text in Gigaword is not morphologically annotated and does not show the diacritics, so diacritics and parts of speech (POS) tagging are added by using, MADAMIRA, which is a part-of-speech tagger that is equipped with text-diacritization feature (Pasha et al. 2014).

The protocol I use to evaluate the stability of the results of the predictive models is 10-fold cross-validations. To do that, the performance of every model is tested 10 times using randomly sampled 10% of the data at each test session. To ensure that the model prediction is tested on every class that exists in the output classes (here, plural patterns), the size of a class that will be predicted by the models has to be equal to or more than number of cross-validations that will be performed. Thus, since the process of cross-validations consists of 10 testing sessions, each class (plural pattern) that will be predicted the model has to occur at least 10 times in the data or has to have at least 10 singular stems. Therefore, a plural pattern that occurs less than 10 times is removed from the data. The data-set after the removal of infrequent plural patterns includes 7533 singular stems. The distribution of the singular stems by plural types is as follows: 2257 feminine sound plurals, 1178 masculine sound plurals and 4098 broken plurals. These singular stems are mapped to 31 plural categories. Figure (5-1) shows these categories and the number of singular stems associated with each category:

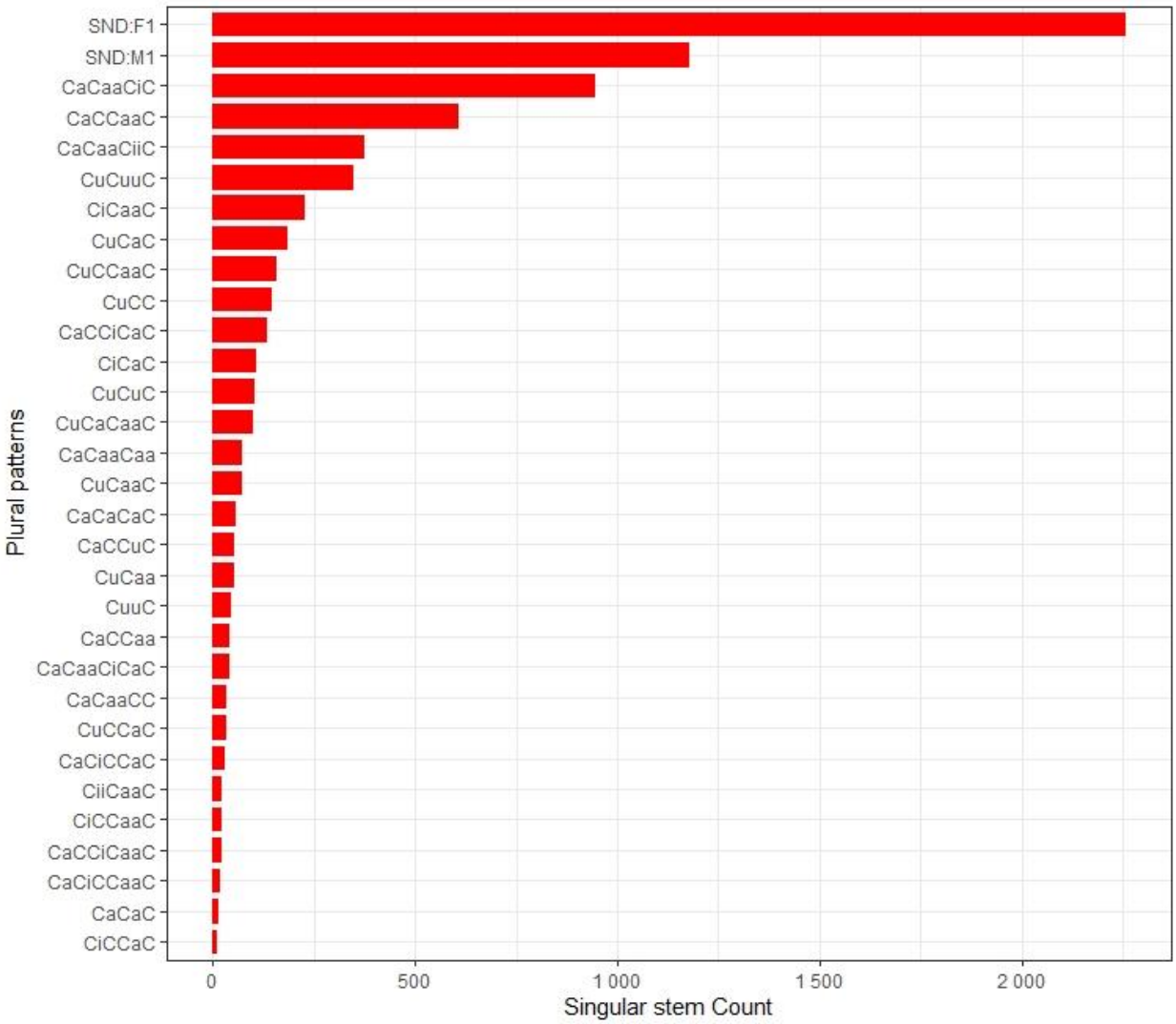


*Figure 5-1*. The plural patterns (y-axis) and the number of singular stems associated with the plural patterns (x-axis).

The kind of computational analysis used in this study requires that input data (the singular stems) to be converted into a feature vector that the model uses to learn how to classify the data. After singular stems are obtained for every plural form, features that will be used to predict plural forms are extracted from all the stems by running a python script written by the author. The study concentrates on the importance of three features of the singular input at plural pattern selection: CV template, Vowel Melody and Human-designation. These features are extracted as follows. First, the human-designation feature is added for each singular stem by the author, such that every singular stem is tagged as Human or non-Human depending on its

meaning. Second, a Python script written by the author converts each singular stem into a CV Template and extracts the Vowel Melody. The three features associated with every singular stem are then added to a feature dictionary to be transformed into feature vector when the computational model is executed. An illustration of the transformation of a singular stem to a feature vector is shown in Figure (5-2):

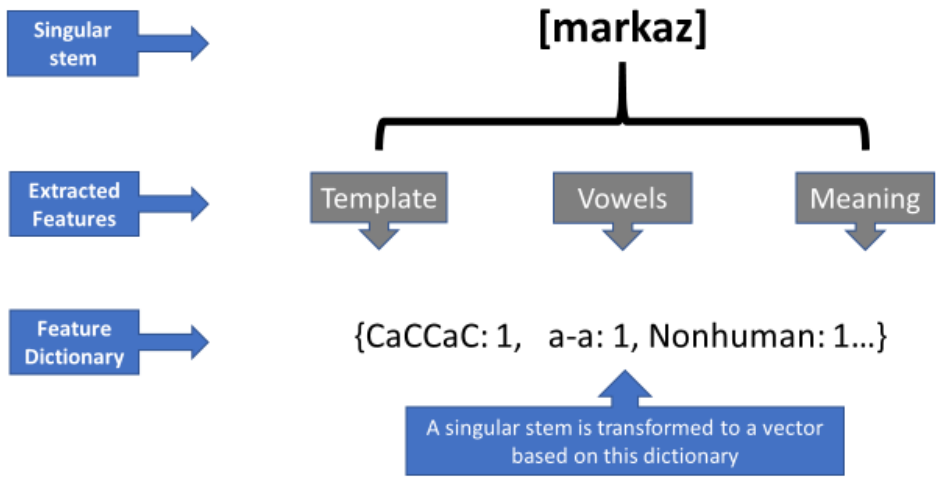


*Figure 5-2.* An illustration of the process of converting the singular stem into a feature vector.

The categories that will be predicted by the models are the vocalized templates of the plural forms. In order to create the templates, plural forms linked with the singular stems are obtained and entered into a Python script that converts the plural forms into vocalized patterns by turning all consonants to C and retaining the vowels. For example, the plural form [maraakiz] for the singular stem [markaz] is turned into a vocalized template by changing all the consonants in that form yielding the pattern *CaCaaCiC*. As indicated in Figure 3-1, there are 31 plural patterns, and each plural pattern represents a class or category that the models have to predict.

### 5.3.2 Model details

In this section, I describe the three K-Nearest Neighbor classifiers I implemented in the model comparison to investigate the importance of the three factors in predictions of plural patterns for given singulars. The first model uses CV template of the singular nouns as a factor to predict their plural patterns. The second model takes vowels and CV templates of the singular stems as factors in plural selection. Third model combines CV template, vowels and semantic quality of human-designation to determine the plural pattern for a given singular.

I implement 10-fold cross-validation, which is a standard method of ensuring that the results of a model are replicable and generalizable (Breiman, Friedman, Olshen, & Stone, 1984). This is performed by iterating the testing procedure for each model 10 times, each time with a randomly sampled 10% of the data-set used as the test group and the remaining 90% used as the training group. The accuracy is the number of times the model selects the correct plural template for a singular stem in the test group.

A model baseline is selected in order to interpret the results. For each test form, the most frequent plural pattern is selected as the baseline.

## 5.4 Results

The accuracy scores from the three KNN models are averaged across the 10 iterations. Table (5-1) shows the accuracy for the three models. A hypothetical model that always predicts the most frequent class in the data is used as the baseline for individual evaluation of the performance of each model. All three models outperform the baseline. Pairwise two-sided dependent t-tests with Bonferroni correction are used to test the difference between all possible combinations of KNN models. The best performing model is the one that uses the template, vowel series and human-designation (70%). It performs significantly better than the model that

uses a feature combination of the template and vowel series, $t(9) = -19.05$, $p < .01$, and better than the model that uses the template only, $t(9) = -35.12$, $p < .01$. There is also a significant difference between the model that uses both the template and the vowel series and the one that uses the template only. The model that uses the template and vowel series (accuracy = 61%) performed significantly better than a model that only uses the template (accuracy = 70%), $t(9) = -16.47$, $p < .01$.

| Model | Accuracy | Standard Deviation |
|---|---|---|
| Template | 0.52 | 0.02 |
| Template, Vowel Series | 0.61 | 0.01 |
| Template, Vowel Series, Human | 0.70 | 0.03 |
| Baseline | 0.30 | - |

*Table 5-1.* Accuracy scores of the KNN models. Scores range from 0 (least accurate) to 1 (most accurate).

To determine the effect of the KNN algorithm on the accuracy of classification (i.e. the dependence of classification on the algorithm), results from KNN models are compared to results from a similar set of Support Vector Machine (henceforth SVM) models. Overall, the SVM models outperformed the KNN models. However, the pattern of results is similar between the two types of algorithm. Table (5-2) shows the accuracy of the SVM models averaged across the 10 iterations. The same set of features used in the KNN models to predict plural patterns are used in the SVM models:

| Model | Accuracy | Standard Deviation |
|---|---|---|
| Template | 0.62 | 0.02 |
| Template, Vowel Series | 0.68 | 0.01 |
| Template, Vowel Series, Human | 0.74 | 0.03 |
| Baseline | 0.30 | - |

*Table 5-2.* Accuracy scores of the SVM models.

As indicated in Table 5-2, the model with the template, vowel series, and human-designating feature performs better than the models with the template and vowel series and the one with the template only. The model with template and vowel series also has higher accuracy than the template-only model. The table also shows that the size of the difference in accuracy between these models is similar to the difference size observed in the KNN models.

Errors made by the 10 implementations of the three sound-and-broken (KNN) models are obtained from confusion matrices and analyzed. Errors are classified into four categories: broken-to-broken errors refer to the type of errors caused by incorrectly classifying a broken pattern as another broken pattern; broken-to-sound errors are errors that occur when the model incorrectly classifies a broken pattern as sound plural; sound-to-broken are caused by confusing a sound plural for a broken plural pattern; and sound-to-sound are the type of errors caused by confusing a sound plural for another class of sound plurals. A breakdown of errors made by the three models is illustrated in Figure 5-3:

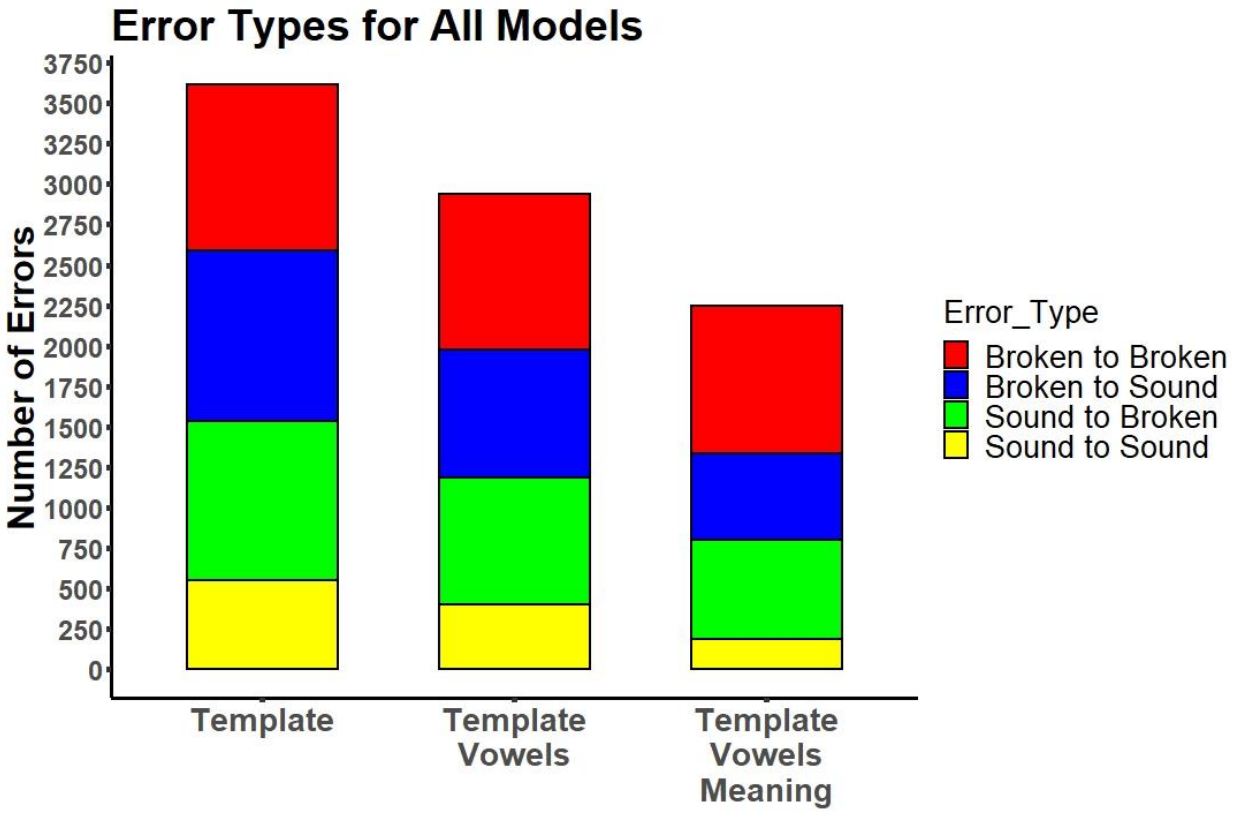


*Figure 5-3.* Error types for all KNN models.

Overall, the error rate decreases as the model uses more features to classify singular stems. The majority of errors in all three models are attributed to confusing a broken plural pattern for another broken plural pattern. Errors caused by incorrectly classifying a broken plural as a sound or vice versa are also frequent. Errors that occur due to confusing one sound plural for another is the least frequent type. The rate of errors in these three classes, namely broken-to-sound, sound-to-broken and sound-to-sound, gradually decreases as the model uses more features to map singular stems to plural patterns, over the three models starting at the model that uses least number of feature (template-only model) to the model that uses the largest number of features (template-vowel-meaning model. The rate of errors that belong to the broken-to-broken type remains roughly similar across the three models.

The performance of the models is the classification of Arabic plural data by plural patterns is also tested using a dataset that includes the broken plural only. Table (5-3) shows the accuracy scores of the models averaged across the 10 iterations when the broken plural patterns are considered only. A baseline that is based on the most frequent category is used to evaluate the individual performance of each model. Overall, all three models perform better than the baseline. Out of the three models compared here, the model that employs template, vowel series and human-designation has the highest accuracy score (72%). This model also differs significantly from the model with template and vowel series, $t(9) = -4.9$, $p < .05$, and the model with template only, $t(9) = -17.62$, $p < .01$. The model that uses template and vowel series (68%) performed better than the model with template only (60%). The difference between the two models is also significant, $t(9) = -10.16$, $p < .01$.

| Model | Accuracy | Standard Deviation |
|---|---|---|
| Template | 0.60 | 0.02 |
| Template, Vowel Series | 0.68 | 0.04 |
| Template, Vowel Series, Human | 0.73 | 0.04 |
| Baseline | 0.23 | -- |

*Table 5-3.* Accuracy of the broken-only KNN models.

The results from KNN models that are applied on the broken plurals are compared to results from a similar set of SVM models. Table (5-4) shows the accuracy of the SVM models averaged across the 10 iterations:

| Model | Accuracy | Standard Deviation |
|---|---|---|
| Template | 0.68 | 0.01 |
| Template, Vowel Series | 0.76 | 0.03 |
| Template, Vowel Series, Human | 0.82 | 0.03 |
| Baseline | 0.23 | -- |

*Table 5-4.* Accuracy of the broken-only SVM models.

The same set of features used in the KNN models to predict plural patterns are used in the SVM models. Overall, the SVM models outperformed the KNN models. However, the pattern of results is similar between the two types of algorithm. An SVM model that is trained on template features (accuracy = 68%) performs better than a KNN model trained on the same features (accuracy = 60%). However, the rate of increase in accuracy as a result of adding vowel and human designation features is similar for two types of models (Template, Vowels: 68% (KNN) vs. 76% (SVM); Template, Vowels, Human: 73% (KNN) vs 82% (SVM)).

## 5.5 General discussion

The goal of this study is to determine the importance of factors such as CV template, vowel series, and human designation in predicting a plural pattern for a given singular. The results of the influence of these factors can be then used to draw conclusions about the generalization and learnability of the plural system in Arabic. The results show that all these features are relevant for the prediction of the shape of the plural template, although with varying degrees. These factors when taken together, successfully predict the plural pattern of the singular stem. The robustness of these features supports the argument that in Arabic nominal pluralization the principal factor in determining the plural form that a particular singular will take is the morphophonological properties of the singular itself.

The results agree with the claim in Levy (1971), McCarthy and Prince (1990) and Ratcliff (1998) that the syllabic shape of the singular forms of Arabic nouns is the major factor in predicting their plural forms. As shown in Table 5-1, a model that is only trained on the template of the singular stem correctly predicts the plural patterns for 52% of unseen singular stems. This claim is based on the notion that singular stems that share a CV template are more likely to be mapped to the same plural pattern. The use of the CV template of the singular noun as a factor to

determine the plural pattern is also consistent with the psycholinguistic research that provides evidence for the use of the template in word processing. In several priming experiments, Boudelaa and Marslen-Wilson (2004) report that word recognition in MSA is facilitated when the prime and the target share the same CV template. The same influence of the CV template on word recognition is also noticed in spoken variety of Arabic (Boudelaa and Marslen-Wilson, 2013).

The results also show that vowel melody of the singular stem contributes to the selection of the plural pattern that the singular will take. As illustrated in *Table 5-1*, adding the vowel series and the CV template improves model accuracy at predicting the plural pattern by 8%. The results agree with the conclusion reported in Levy (1971) Ratcliff (1998) that considers singular stem vocalism as a secondary factor to predicting the plural form for a given singular stem. These results, however, are inconsistent with those reported in psycholinguistic experiments that fail to find evidence for the effect of vowels on word processing. Using a priming experimental design, Boudelaa and Marslen-Wilson (2004) compared the priming effect on lexical decisions when the prime and the target shared the CV template versus when they share a CV template and vowel melody. They found that the amount of priming induced by a shared CV template did not significantly differ from the amount of priming induced by a shared CV template and vowel melody.

In addition to the template and vowels, the results find that semantic properties of the singular stem, namely whether the singular designates human beings, plays a role on determining the plural pattern of a given singular stem. Adding the semantic feature to a model that uses a CV template plus vowels improves its accuracy by 9%. This is in line with the results reported in

Levy (1971) that found a systematic relationship between the singular stem rationality and the type of the plural pattern it will take.

The analysis of errors made by the three KNN models reveals that the majority of errors stem from confusing one broken plural pattern for another (broken-to-broken errors). Errors attributed to incorrectly selecting a broken plural for a word that takes a sound plural or vice versa also exist, but this type of error decreases as the model uses more factors. Unlike these errors, the error rate attributed to selecting an incorrect broken plural pattern for a word that takes a particular broken plural remains stable across the three models. One possible explanation for this pattern of errors is that, in Arabic broken plurals, multiple plural patterns can be mapped to singular stems that share the same prosodic shape. As shown in 4.5, plural patterns within a particular weight class are sometimes linked to singulars of the same syllabic shape. This crossover between singular stems that share syllabic shape may results in an increase in broken-to-broken type of errors when a model attempts to predict the plural pattern for such stems.

Each model trained on the broken-only data (Table 5-3) outperforms the model that uses the same set of factors but, at the same time, includes broken and sound plural data (Table 5-1). However, the extent of the influence of these factors varies between the two types of models (broken-only vs. sound-and-broken). The results show that the extent of the influence of template and vowel features on the overall performance is comparable between the model that is implemented on sound and broken plurals and the model that is restricted to the broken plurals. For example, the accuracy of the two models (i.e. sound-and-broken and broken-only) improves by 8% when they are trained on the template and vowel features. However, differences in the effect of the features emerges when the feature of human designation is considered. The results show that using a combination of template, vowels, and human-designation features improves

the accuracy of the model implemented on sound and broken plurals by 10%, whereas using the same set of factors to train a model implemented on only the broken plural data-set improves its accuracy by 5%. The difference in the influence of human-designating property between the model with broken and sound plurals and the model that is restricted to the broken plurals indicates that the designation of human entities plays an integral role in predicting the plural pattern for singular stems that take sound plurals. These findings lean toward Levy's (1971) conclusion that the overwhelming majority of Arabic nouns that take sound plurals tend to be those that denote human entities where the nouns that take broken plurals include human and non-human nouns.

The computational study indicates that although the nominal plural system in Arabic is complex and characterized by a many-to-many mapping, it is possible to predict plural patterns on the basis of the semantic and morphophonological features of the singular stems. Computational models are able to make generalizations to unseen nouns based on training on these features. However, the question of how much of this generalization is attributed to the semantic and morphophonological features and how much to the learning algorithm needs further investigation. Partitioning the information from the models into parts that are explained by the semantic and morphophonological features and parts that are attributed by the algorithm is an important question for future research.

One of the limitations of the current study is that it examines the effect of the semantic properties represented by focusing on just one semantic feature, namely the property of human designation by the singular nouns. According to Yaaqub (2004), Arab grammarians list semantic properties such as abstractness and agentivity among the factors that influence the mapping of singular nouns to their plural forms. Levy (1971) in a statistical investigation of the distribution

of singular and plural forms shows that these semantic properties play a moderate role in the selection of particular broken plural patterns. The current data-set is not marked for abstractness and agentiveness, however. Testing the effect of these semantic properties on the prediction of plural pattern is an opportunity for future research.

Another opportunity for future research is to conduct psycholinguistic experiments to evaluate the importance of these cues on the selection of plural patterns by Arabic speakers. This would provide a chance to contrast the results from the computational analysis with the behavior of native speakers in learning and making morphological generalizations.

## Chapter 6 Conclusion and Future Direction

The dissertation investigates three problems in Arabic plural system: the distribution of sound and broken plurals, the role of stem weight on the singular-to-plural mapping, and the factors that contribute to this mapping. In this section, I will give a summary of the results from each study. Then, I will try to address research limitations and outline the opportunities for future research.

The results from the statistical distribution of Arabic sound and broken plural types show that the distribution of broken and sound plurals in Arabic is in line with the description of Arabic pluralization as a minority default system, one where the regular rule-based morphological operation becomes less frequent than the irregular one. However, the difference between the count of noun types that take sound plurals and those that take broken plurals is marginal. The results also show that token frequency which represents the number of times a plural noun appears in actual language use differs significantly between the two types of plurals. In spite of the fact that the difference between the count of the noun types taking sound plurals and that of the noun types taking broken plurals is marginal, nouns that take broken plural patterns are used more frequently than nouns that take sound plurals. These results are consistent with the usage-based model of morphology (Bybee 2001) which predicts the tendency of irregular patterns to have higher token frequency to block regularization.

The results also provide evidence for the role of stem weight as described in Heath (2005, 2018) on the broken plural patterns in Arabic. Broken Plural patterns can be effectively grouped

into coherent classes based on the number of segments in their singular stems. The use of stem weight to account for the formal relations between singular nouns and their plural forms in Arabic positions Arabic with languages such as Tamashek, and Penange that show forms of morphological alternations that are sensitive to weight of the stem inputs.

The results of the computational predictive models show that features of Arabic singular nouns, such as CV template, vowel melody and human designation, are all instrumental in the selection of the plural patterns. These features, however, differ in their importance in predicting the plural pattern for a given singular noun. The CV template carries most of the predictive power, followed by the vowel melody and the semantic property of human designation. The analysis of the error types reveals that broken plurals account for the majority of the errors made by the models. These errors result from confusing one broken plural pattern for another. Errors attributed to incorrectly selecting a broken plural for a word that takes a sound plural or vice versa also exist, but this type of error decreases as the model uses more factors. Unlike these errors, the error rate attributed to selecting an incorrect broken plural pattern for a word that takes a particular broken plural remains stable across the three models.

The three studies in this dissertation give a comprehensive view of the increased complexity and ambiguity featured in Arabic nominal plurals. They also emphasize the role of morphophonological and morphosemantic features of both singular and plural nouns in the explanation of frequency trends, mapping relationship and morphological generalizations. It is almost impossible to examine problems related to the Arabic plural system without the involvement of these features. While many models of Arabic morphology rely on aspects of phonology and morphology, none of the current models explicitly incorporates semantics in its design, in spite of the essential role these features have. One of the major findings of this

dissertation is recognition of the contribution of semantic features in the analysis of Arabic pluralization. Therefore, a comprehensive model of non-concatenative nature of Arabic plurals ought to include aspects of semantic properties of the word, in addition to its morphophonological properties.

One of the limitations of these studies is related to the language variety on which the analysis is conducted. The analysis only covers one variety of Arabic, specifically MSA. There is a sociolinguistic dispute on whether MSA is representative of the linguistic behavior of Arabic speakers. Given the diglossic nature of Arabic (Ferguson 1959), MSA as the variety for formal uses would differ from the regional dialects that Arabic speakers communicate with in natural informal settings. A possible opportunity for future research is to conduct a similar computational and qualitative analysis of sound and broken plurals in different spoken varieties of Arabic. This will address the sociolinguistic concerns about the use of MSA as representative of Arabic speakers and will also provide a chance to contrast the plural system in MSA with that of other varieties.

Another limitation of the dissertation is related to the lack of insights from psycholinguistic research on the reality of the findings. The findings that the dissertation have reached are promising and can be beneficial on a descriptive and pedagogical level. However, if the concern is to make claims about processing and learnability of Arabic plurals, then it is important to rely on psycholinguistic experiments to determine the psychological reality of such claims. Verifying the findings reached in this dissertation through vigorous experimental methods and tools in psycholinguistics is another opportunity for future research.

**Bibliography**


Al-Sulaiti, L. (2009). *Corpus of Contemporary Arabic [dataset]*. Retrieved August 25, 2009, from http://www.comp.leeds.ac.uk/eric/latifa/CCA_raw_utf8.txt.

Abu Al-Saud, Abbaas (1971). *Al-Faysal fi Alwan Al-Jumuuʕ* [The Decisive Criterion for Types of Plurals]. Cairo: Dar Al-Macarif.

Boudelaa, Sami & M. Gaskell (2002). A re-examination of the default system for Arabic plurals. *Language and Cognitive Processes*, 17, 321–343.

Boudelaa, S., & William Marslen-Wilson (2004). Abstract morphemes and lexical representation: The CV-skeleton in Arabic. *Cognition*, 92, 271–303.

Boudelaa, S., & William Marslen-Wilson (2004). Morphological structure in the Arabic mental lexicon: Parallels between standard and dialectal Arabic. *Language and Cognitive process*, 28:10, 1453-1473.

Brame, Michael (1970). *Arabic phonology: implications for phonological theory and historical Semitic.* Doctoral dissertation, MIT.

Breiman, Leo, Friedman, Jerome, Olshen, Richard, & Charles Stone (1984). *Classification and regression trees.* Belmont, CA: Wadsworth.

Buckwalter, Tim (1997). Issues in Arabic morphological analysis. In A. Soudi, A. van den Bosch, & G. Neumann (Eds.), *Arabic computational morphology: Knowledge-based and empirical methods.* (vol. 38, pp. 23–41). Dordrecht: Springer.

Bybee, Joan (2001). Phonology and Language Use. Cambridge: University Press.

Cattell, Raymond B. (1966b). The scree test for the number of factors. *Multivariate Behavioral Research*, 1, 245–276.

Clements, George N. & Samuel J. Keyser (1983). *CV Phonology: A Generative Theory of the Syllable*. Cambridge, Mass: MIT Press.

Comrie, Bernard, (1991). On the importance of Arabic to general linguistic theory. In: B. Comrie and M. Eid (eds). *Perspectives on Arabic Linguistics: Papers from the Annual Symposium on Arabic Linguistics III*. Amsterdam: John Benjamins, 3–30.

Dawdy-Hesterberg, Lisa G. & Janet Pierrehumbert (2014) Learnability and generalisation of Arabic broken plural nouns. *Language, Cognition and Neuroscience*, Vol. 29, 10, 1268-1282.

Ding, Chris & Xiaofeng He (2004) Proceedings of the 21st International Conference on Machine Learning, 2004.

Ferguson, Charles A. (1959). Diglossia. Word 15:2, 325-340.

Frisch, Stefan, Janet Pierrehumbert, and Michael Broe (2004). Similarity Avoidance and the OCP. *Natural Language and Linguistic Theory*, 22, 179-228.

Goldsmith, John (1976). *Autosegmental Phonology*. Bloomington: Indiana University Linguistics Club.

Hammond, Michael (1988) Templatic transfer in Arabic broken plurals. *Natural Language and Linguistic Theory*, 6, 247-270.

Hayes, Bruce (1989). Compensatory lengthening in moraic phonology. *Linguistic Inquiry*, 20, 253-306.

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

Heath, Jeffrey (2005). *A Grammar of Tamashek: Tuareg of Mali*. Mouton de Gruyter: New York.

Heath, Jeffrey (2018). *A Grammar of Penange (Dogon, Mali)*. Language description heritage library, University of Michigan.

Holes, Clive (2004). Modern Arabic: Structures, Functions, and Varieties. Washington, DC: Georgetown University Press.

Hyman, Larry (1985). *A Theory of Phonological Weight*. Dordrecht: Foris.

Jolliffe, I. T. (1986). Principal component analysis. New York: Springer.

Kaiser, Henry F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement,* 20, 141–151.

Khouloughli, Djamel E. (1992). Basic Lexicon of Modern Standard Arabic. L'Harmattan, Paris.

Lane, Edward W. (1863) An Arabic-English Lexicon 8 vols., London: Williams & Norgate.

Levy, Mary M. (1971) *The Plural of the Noun in Modern Standard Arabic*. Doctoral Dissertation, University of Michigan, Ann Arbor.

Maddieson, Ian (1984) Patterns of Sounds. London: Cambridge University Press.

McCarthy, John (1979). *Formal Problems in Semitic Phonology and Morphology*. Doctoral dissertation, MIT.

McCarthy, John & Alan Prince (1990). Foot and words in prosodic morphology: the Arabic broken plural. *Natural Language and Linguistic Theory*. 8, 209-284.

Murtonen, Aimo E. (1964) Broken Plurals, the Origin and Development of the System. Leiden: E. J. Brill.

Nakisa, Ramin, Plunkett, Kim & Ulrike Hahn (2001). Single- and dual- route models of inflectional morphology. In P. Broeder & J. Murre (Eds.), *Models of language acquisition: Inductive and deductive approaches* (pp. 201–222). Cambridge, MA: MIT Press.

Newman, Paul (2017) Syllable weight as a phonological variable. *Syllable weight in African languages*, ed. by Paul Newman 9 -24. Benjamins: Philadelphia.

Nosofsky, Robert (1990). Relations between exemplar-similarity and likelihood models of classification. *Journal of Mathematical Psychology*, 34, 393–418.

Parker, R., Graff, D., Chen, K., Kong, J. and Maeda, K. (2011) *Arabic Gigaword*. LDC Catalog No.: LDC2011T11. Linguistic Data Consortium.

Pasha, A., Al-Badrashiny, M., Diab, M.T., Kholy, A.E., Eskander, R., Habash, N., Pooleery, M., Rambow, O., & Roth, R. (2014). MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. *LREC* 14, 1094-1101.

Pierrehumbert, Janet B. (1993). Dissimilarity in the Arabic Verbal Roots. in *Proceedings of the North East Linguistics Society*, 23, 367–381.

Pinker, Steven, and Alan Prince. 1988. On Language and Connectionism: Analysis of a Parallel Distributed Processing Model of Language Acquisition. *Cognition*, 23, 73-193.

Plunkett, Kim and Ramin C. Nakisa (1997) A connectionist model of the Arabic broken plural system. *Language and Cognitive Processing*,12 (5/6), 807 - 836.

Prince, Alan & Paul Smolensky (1993) *Optimality Theory: constraint interaction in generative grammar*. Manuscript, Rutgers University, New Brunswick, and University of Colorado, Boulder.

Ratcliffe, Robert (1990) Arabic broken plurals: Arguments for a twofold classification of morphology. In M. Eid and J. McCarthy (eds). *Perspectives on Arabic Linguistics II*. Amsterdam: John Benjamins, 94-119.

Ratcliffe, Robert (1998) *The 'Broken' Plural Problem in Arabic and Comparative Semitic*. Amsterdam: John Benjamins.

Rumelhart, David & James McClelland (1986). On learning the past tense of English verbs. In D. Rumelhart & J. McClelland (eds). *Parallel distributed Processing*, Vol. 2. Cambridge MA: MIT Press.

Versteegh, Kee (1997). The Arabic Language. New York: Columbia University Press.

Watson, Janet (2002) The Phonology and Morphology of Arabic. Oxford: Oxford University Press.

Wehr, H. (1976). *Arabic–English dictionary: The Hans Wehr dictionary of modern written Arabic* (3rd ed.). Ithaca, NY: Spoken Language Services.

Wright, W. (1971). *A grammar of the Arabic language*. Cambridge: Cambridge University Press.

Yaaqub, Emile B. (2004) *Al-Muʕjam Al-Mufassal fi Al-Jumuuʕ* [The Detailed Dictionary for Plurals]. Beirut: Lebanon.

Youmans, Gilbert (1990) Measuring lexical style and competence: the type-token vocabulary curve'. *Style*, 24, 584-599.