NORTHWESTERN UNIVERSITY

The Structural and Statistical Basis of Morphological Generalization in Arabic

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Linguistics

By

Lisa Garnand Dawdy-Hesterberg

EVANSTON, ILLINOIS

December 2014

ABSTRACT

The Structural and Statistical Basis of Morphological Generalization in Arabic

Lisa Garnand Dawdy-Hesterberg

This thesis examines learnability and generalization of the morphology of Modern Standard Arabic, focusing on two subsystems: the noun plural and the masdar of form I verbs. Using psycholinguistic experiments and computational analyses, I assess two aspects of generalization. First, I address the learnability of a morphological system based on the predictability of the morphological variant of an unseen form based on analogy to existing forms. Second, I assess how speakers generalize existing morphological patterns to previously unseen forms using nonce-form tasks.

More generally, this thesis investigates how speakers learn and generalize morphological patterns in systems with two characteristics: coarse-grained representations, as both systems contain non-concatenative patterns that require a high level of abstraction to represent; and high uncertainty, in that there are 30+ patterns for both systems under investigation, and the pattern that an existing words takes is somewhat unpredictable. By studying systems with these characteristics, I examine the key questions of: 1) what is the basis of analogy in morphological generalization in Arabic?; and 2) how do speakers decide among the possible outcomes when there are a large number of possibilities with varying likelihoods?

First, I demonstrate that speakers generalize existing noun plurals primarily on the basis of the coarse-grained CV template, and select among the possible morphological variants in a probabilistic manner, indicating that they track lexical statistics on this coarse-grained level. Second, I demonstrate that the masdar is quite predictable on the basis of type statistics on the

coarse-grained representation of the verb pattern, which disproves previous claims that the masdar of form I verbs is unpredictable. Finally, in a nonce-form task, I show that speakers also generalize existing masdar patterns in a probabilistic manner, but do so not on type statistics on the verb pattern, but potentially on the CV template. For both of these systems, speakers utilize the full range of possibilities in generalization, indicating that they generalize low-probability patterns even when there are 30+ possibilities. The implications of these findings for theories of Arabic morphology as well as theories of learnability and generalization of complex morphological systems are discussed.

ACKNOWLEDGEMENTS

First and foremost, a great deal of thanks goes to my committee members, without whom this dissertation would never have been possible. Thanks first to Doug Downey, whose Machine Learning class first sparked a deeper interest in using computational methodologies in my research, and who, despite coming from a very different field and background than myself, was always willing and happy to discuss new approaches to issues I encountered along the way. Thanks to Matt Goldrick, who throughout my years at Northwestern has been an amazing teacher, mentor, and cheerleader. Beyond his help with methodological issues and finding new ways to think about my data, he has been an enormous source of support in academic and non-academic issues alike. Finally, and most importantly, thanks to my chair, Janet Pierrehumbert, who has inspired me in many ways to be a better scientist. She has always encouraged me when things became difficult, and I've learned an enormous amount from her about how to think critically, frame my questions in the best way, and to look beyond my field for useful sources and methodologies. I owe her a debt of gratitude for all of her help and support in finishing this thesis, and in becoming a better scientist.

A number of other people and organizations have also been extremely helpful and supportive in this research. Thanks to Chun Liang Chan for technical support in creating my web-based experiments, and for developing the back-end support for them. Thanks to Yaseen Mansour, my undergraduate RA, both for help in creating stimuli and for stimulating discussions about issues in Arabic. The Language Dynamics Lab Group has been a source of feedback throughout this process, and I thank all of the members for their advice and help on this research. Thanks go to the Graduate Grants Committee at Northwestern University for financial support for my research. Finally, thanks to the Wordovators grant from the John Templeton Foundation.

This project was made possible through the support of a grant from the John Templeton Foundation. The opinions expressed in this publication are those of the author and do not necessarily reflect the views of the John Templeton Foundation.

My time at Northwestern would not have been the same without a number of people. First, my cohort, Jordana Heller, Jenna Luque, Charlotte Vaughn, Kyounghee Lee, and Elizabeth Mazzocco, were all instrumental in surviving my first few years here. I'll always have fond memories of the many late nights spent working on assignments in Swift 007. Thanks especially to Jenna Luque, who has been a more supportive and encouraging friend than I could ask for. To Jason French, thanks for always being up for a coffee run, and most of all, for listening.

Finally, my family deserves acknowledgement for being so wonderful and supportive, both during grad school and throughout my life. My mother has always been there for me, and my biggest cheerleader. My sisters, Kirstin and Erika, are both among my closest friends, and I wouldn't wish it any other way. My father first inspired me to pursue a PhD, and has been an inspiration to me since I was a child. My grandparents, Laboyta and Richard, have also been part of the most supportive family I could hope for. Finally, my husband, Adam, has always been my best friend and my biggest supporter, and I couldn't imagine life without him. Here's to our future together, post-dissertation.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

Chapter 1 : Introduction

1.1 What is morphological generalization?

In word-formation, speakers must integrate a great deal of linguistic information in order to create a word that is both appropriate to their needs and comprehensible to listeners. During this process, speakers draw on stored information from the lexicon to guide their selection and application of a morphological pattern. For example, a speaker might coin a word "turquoisify" if the speaker dyed his shirt turquoise and wished to describe this process in a novel way. The speaker might elect to use the suffix [-ify] with "turquoise" on the basis of existing words with the suffix [-ify], such as "gentrify" and "acidify." There may be competition from affixes that occupy the same semantic field, such as [-en] in "redden" or "darken" and [-ize] in "personalize" or "industrialize."

There are many factors that influence the selection of [-ify] in this case, including the phonological and phonetic features of the base and affix, frequency of existing forms and patterns in the lexicon, and the similarity of the new form to existing forms in the lexicon (e.g., Albright & Hayes, 2003; Ernestus & Baayen, 2003; Hay, 2003; Hay & Baayen, 2005). Before a morphological process can be applied to new forms, it must be generalized at some level. This could occur through rule formation, or on-line abstraction across stored forms (e.g., Albright & Hayes, 2003; Bybee, 1995; Ernestus & Baayen, 2003; Prasada & Pinker, 1993; Rumelhart & McClelland, 1986; Stemberger & MacWhinney, 1988). Further, the selection of the specific morphological process or affix to use is subject to uncertainty, as speakers often have a number of choices of the process or affix to use, as in the above example, where there are at least three suffixes that could be used to form this new word. Word-formation is thus an avenue for studying generalization of linguistic patterns within a complex decision space.

The morphology of Modern Standard Arabic (henceforth Arabic) is both an intriguing and important venue to study the processes of word formation for a number of reasons. First, the morphology is complex, in that the many non-concatenative morphological processes make significant structural changes to the base form. In addition, there are many possible patterns for some morphological processes. This system is ripe for studying generalization, as speakers must abstract complex morphophonological patterns in order to generalize to new forms, and the decision space for deciding on the appropriate pattern to apply is large.

Critically, Arabic morphology provides an ideal case study with which to examine a number of unsolved issues in morphological processing and word-formation. As noted above, there are often a large number of linguistic factors that affect the selection of an affix or process with which to form a new word. In order for a speaker to learn a morphological process, they must be able to generalize it in some way. The way in which speakers generalize and represent complex non-concatenative morphological processes gives insight into the linguistic features that speakers attend to in learning these processes. Second, the manner in which speakers select among morphological processes when there are many available patterns varies in the literature depending on the complexity of the system, particularly in terms of the relative probabilities of the possible variants. There has been little examination of systems where there are a comparatively huge number of morphological variants within a given system (30+ in some Arabic systems), and how this large decision space interacts with the likelihood of a particular choice. The examination of morphological systems with the factors of non-concatenative morphology, and high uncertainty in terms of both a large number of patterns and the relative unpredictability of the optimal pattern to apply to a given form, gives insight into the interaction of these factors in learnability and generalization in complex morphological systems.

First, I will explain what non-concatenative morphology is, and give a general overview of Arabic non-concatenative morphological processes. I will then outline the specific morphological subsystems that this thesis examines. Next, I will examine analogy formation in morphological generalization, and how the study of this process can give insight into learning and generalization of complex morphological representations and systems. Finally, I will give a brief outline of the three body chapters and the experiments and computational analyses therein.

1.2 Arabic non-concatenative morphology

The study of Arabic morphology allows us to examine how speakers cope with high uncertainty and a large decision space in learning and generalization. In addition, the complex non-concatenative morphological structure gives insight into how speakers learn abstract morphological representations. In both of the sub-systems examined in this thesis, there are a large number of possible morphological patterns that can apply to a given word, with as many as 33 possible patterns cited for the noun plural (McCarthy & Prince, 1990a; Wright, 1988), and as many as 44 cited for the masdar (Wright, 1988). For both of these systems, most or all of the morphological patterns are non-concatenative, which adds a degree of complexity to learning.

First, I will explain what non-concatenative morphological structure is, and why this is important and under-examined in the domain of learnability. In English and other language with concatenative morphology, words are generally formed linearly by adding suffixes and prefixes to existing stems or words, as in [pre] + [determine] to form *predetermine* or [happy] + [ness] to form *happiness*. In such concatenative morphological systems, most roots can appear in isolation in surface forms, as in the examples above, although some stems only appear across derived

word forms, for example [duce] as in [re] + [duce] to form *reduce* and [in] + [duce] to form *induce*. In languages with nonconcatenative morphology, in contrast, words are formed with interleaved morphemes that do not generally appear in isolation in surface forms (McCarthy & Prince, 1990a; Wright, 1988). For example, in Figure 1.1, under the framework of Prosodic Morphology (McCarthy, 1981, 1982; McCarthy & Prince, 1990b), there are four morphemes in the deverbal place noun [madrasaT] "school." The **CV template** specifies the prosodic and skeletal structures of the word, specifying whether each phoneme is a consonant or a vowel. This tier has been shown to be the dominant factor in various morphological processes in Arabic, including noun pluralization, diminutivization and verb measure derivation (Dawdy-Hesterberg & Pierrehumbert, 2014; McCarthy, 1981, 1993; McCarthy & Prince, 1990a) The **vocalic melody** specifies the vowels in the V slots in the CV template. As shown in Figure 1.1, the vocalic melody exhibits rightward spreading if there are fewer vowels specified in this tier than the number of V-slots in the CV template. The **x-morpheme** contains any consonants other than those in the **verbal root**. In the case of the example below, the **x-morpheme** contains only one element, [m], which is frequent in deverbal place nouns. The fourth and final tier is the **verbal root**, which contains the consonants that specify the remaining C-slots in the CV template. In general, verbal roots are tri-consonantal, and they largely specify semantic information (Prunet, 2006). As with the vocalic melody, the **verbal root** exhibits rightward spreading if fewer consonants are specified in the tier than the number of C-slots, for example as in [ħabiib] ("lover"), the verbal root of which is the bi-consonantal [ħ b]. The non-linear combination of morphemes in non-concatenative morphology necessitates abstract representation of the morphemes at some level, which adds a layer of complexity to learning such morphological systems.

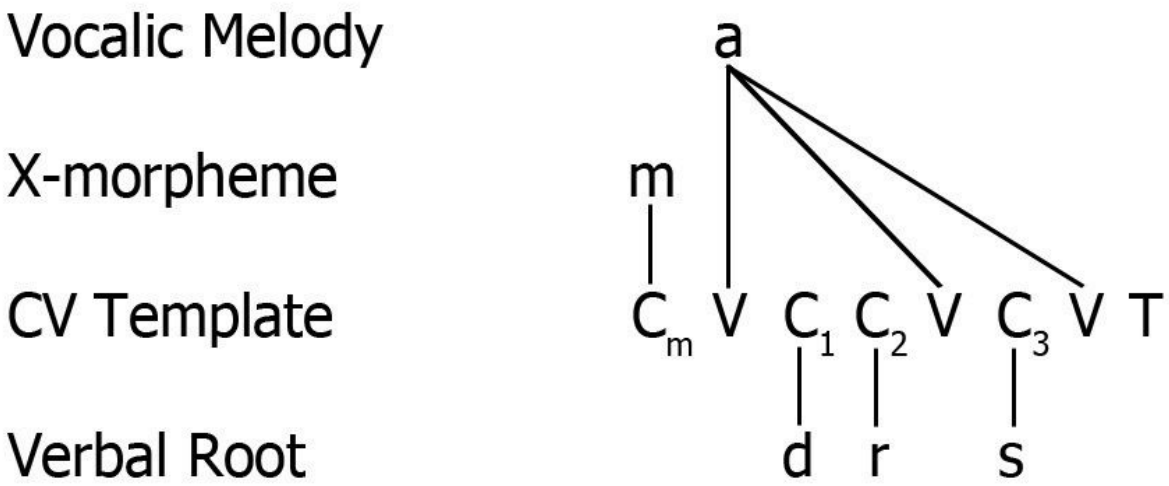

Figure 1.1: Tier structure of [madrasaT] "school" (from Dawdy-Hesterberg & Pierrehumbert, 2014)

In combination, the three non-verbal root tiers (CV template, vocalic melody, and x-morpheme) form the **pattern**. Although it is not generally considered a morpheme, it captures common derivational and inflectional paradigms. In addition, the pattern specifies much of the grammatical information for the word, such as word type, gender and animacy, as well as some semantic information, which are not necessarily specified by the tiers independently. In Figure 1.1, the pattern is [ma$C_1C_2aC_3$aT], and this pattern specifies that the word is a noun, feminine, singular and inanimate. The combination of the pattern and the verbal root is somewhat combinational in terms of the semantics of the word. The pattern [ma$C_1C_2aC_3$aT] generally means "place associated with X," while the verbal root [d r s] denotes "study", so the resulting word means, roughly, "place associated with study," or more specifically, "school" in standard usage. Although word meanings are not purely combinational due to a variety of factors, such as

semantic drift, this example illustrates the general roles of the verbal root and pattern in negotiating word meaning.

The above example shows a word formed via a nonconcatenative morphological process, the formation of which is complex for a variety of additional reasons. As noted above, in concatenative morphological processes, the morphemes are generally linearly separable, and the changes to the base form are minimal and typically occur at the morpheme boundary. For example, in [electricity] from [electric] + [ity], there are two regular and generally predictable phonological changes: the [k] in [electric] becomes [s], and the stress shifts to the antepenultimate syllable of the derived form. In non-concatenative morphological processes, the internal structure of the base form changes significantly, and there are often a number of additional phonological changes. Moreover, these changes are non-local, in that they are spread across the entire word. Figure 1.2 below shows the singular and plural form of "student" in Arabic, which takes a non-concatenative ("broken") plural pattern. In this example, there are three major structural changes from the singular to the plural. First, the vowel in the first syllable is shortened. Second, the vowel in the second syllable is lengthened. Third, the second verbal root consonant is geminated. These structural changes can all be captured by a change in CV template, but to a learner, it may not be the case that these changes are obviously linked. In addition to the structural changes, there are segmental changes via a change in the vocalic melody from [a a i] to [u a]. In sum, the large number of non-local structural and segmental changes make nonconcatenative morphological processes difficult for learners to acquire.

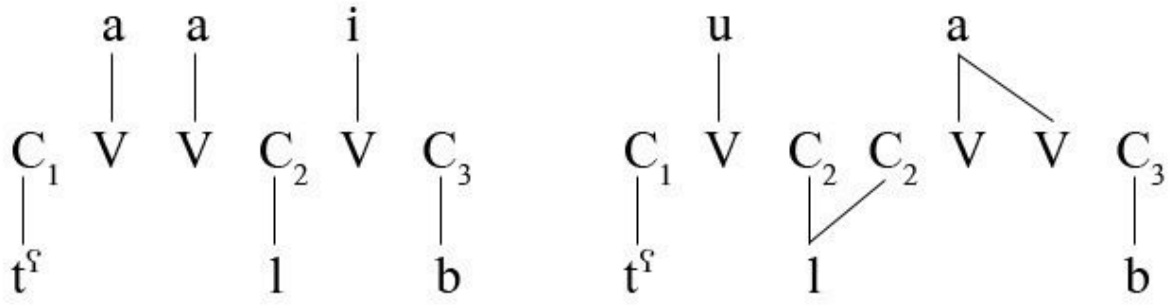

Figure 1.2: Tier structures of singular and plural "student" (from Dawdy-Hesterberg & Pierrehumbert, 2014)

As noted, Arabic morphology is difficult for the learner for two main reasons: 1) there are a large number of possible patterns that a given base form could take in inflection/derivation, and 2) the non-concatenative morphology requires abstract representation of the morphemes to capture the non-local changes a form undergoes. In addition to the large number of possible patterns, an additional source of complexity is that a large percentage of forms take patterns other than the dominant one. In the noun plural system, as noted above, there are as many as 33 possible patterns, of which 31 are non-concatenative (McCarthy & Prince, 1990a; Wright, 1988). The proportion of non-dominant patterns is high, estimated to be 26-41% in two corpus studies (Boudelaa & Gaskell, 2002; Dawdy-Hesterberg & Pierrehumbert, 2014). In comparison, 2% of English noun plural types and 14% of English past tense verb types take non-dominant (also frequently called 'irregular') patterns (Marcus, Brinkmann, Clahsen, Wiese, & Pinker, 1995). Thus, in order to generalize an existing pattern to a new form, speakers must cope with both the abstract morphological structure and the large number of possible applicable patterns, some of which are relatively frequent.

This thesis focuses on an analogical approach to morphological generalization. Thus, the focus will be on how speakers select a morphological pattern from among the existing patterns in a system, without specific reference to how the pattern is applied to the base form. The rules or

schemas for applying a morphological process to a base form can certainly be derived from such an approach, but this thesis does not focus on creating these rules or schemas. Instead, this thesis seeks to explain how speakers learn and generalize morphological patterns given the distribution and structure of existing words in the system.

This thesis will examine two specific subsystems, both of which exhibit the relevant properties of having largely non-concatenative morphology, and a large number of possible patterns. The noun plural, as mentioned, is relatively well studied. There is both theoretical and modeling work indicating that the CV template plays the primary role in plural selection (Dawdy-Hesterberg & Pierrehumbert, 2014; McCarthy & Prince, 1990a). However, the modeling work has achieved only 65-70% accuracy in predicting plural patterns for unseen forms based on existing forms, where a random type-frequency-weighted baseline achieves 39% accuracy (Dawdy-Hesterberg & Pierrehumbert, 2014; Nakisa, Plunkett, & Hahn, 2001; Plunkett & Nakisa, 1997). Additionally, there has been no systematic study of speaker behavior in pluralization of nonce forms, which these studies seek to address.

The verbal noun ("masdar") of form I (underived) verbs is less studied and potentially more unpredictable than the noun plural. Wright (1988) lists 44 possible non-concatenative masdar patterns for form I verbs, although 15 of these are indicated as rare. A brief survey of 188 common regular verbs from (Holes, 2004), a popular Arabic textbook (Brustaad, Al-Batal, & Al-Tonsi, 2004), and a dictionary of verbs for Arabic learners (Scheindlin, 2007) yields 23 masdar patterns. The most frequent pattern [CaCC] comprises 38% of the set, with the two next most frequent [CaCaC] and [CuCuuC] comprising 12% each. It is unclear if the statistically dominant pattern is the default in this system, and, to my knowledge, there is no existing account of the factors that drive masdar formation for form I verbs. Further, there is no study to my knowledge

of the extent to which native speakers of Arabic know this system, and can generalize existing patterns to new forms.

These two systems are of particular interest both because of the large number of possible patterns, and also because the proportion of non-default (or 'irregular') patterns is high. Additionally, there is some work on the statistical basis of analogy for the noun plural, but there is no such work for the masdar. It is unclear if similar factors govern morphological generalization for both systems, and for Arabic morphology on the whole. The main goal of this research is to better understand generalization of morphological patterns in non-concatenative morphology via two main approaches: analogical modeling of morphological generalization, and psycholinguistic examination of speaker behavior in word formation using nonce-form generalization tasks.

1.3 Definitions of key terminology

CV Template: The CV template is the skeletal structure of the word as defined in McCarthy (1981, 1982, 1993), and McCarthy and Prince (1990a). This abstract morphological tier is the base to which the other morphological tiers are associated (see Figure 1.1). In traditional accounts of Arabic morphology, the CV template contains only the segments **C** and **V**, which represent consonants and vowels, respectively (Hammond, 1988; McCarthy & Prince, 1990a). In the current account, following Dawdy-Hesterberg and Pierrehumbert (2014), the CV template contains an additional segment **T**, which represents the feminine grammatical suffix *taa marbuta* ("bound t"). This segment is considered a separate part of the CV template from other consonants in Dawdy-Hesterberg & Pierrehumbert for two reasons. First, forms with the suffix behave

differently in morphological processes than unsuffixed forms (Dawdy-Hesterberg & Pierrehumbert, 2014; McCarthy & Prince, 1990a; Ratcliffe, 1998), and suffixed forms need not be derived directly from their unsuffixed counterparts. Second, unlike other consonants in the template, *taa marbuta* has variable behavior in syllable weight, which is specific to this grammatical suffix.

Pattern: The pattern refers to the partially-filled CV template, where all segments in the vocalic melody and x-morpheme tiers are specified but the verbal root is not. For example, the pattern for [ʔawlaad] "boys" is [ʔaC1C2aaC3]. The pattern carries much of the grammatical specification of the word, which is not independently tied to the constituent parts (Prunet, 2006).

Analogy: Many theories of morphological productivity have argued that generalization to new forms occurs on the basis of analogy to existing forms, such that new forms take the morphological pattern of the most similar existing form(s) (e.g., Prasada & Pinker, 1993; Rumelhart & McClelland, 1986; Skousen, 1989, 1993). The critical element examined here is the basis of this similarity in Arabic, and the extent to which this definition of similarity varies for morphological systems with differing morphophonological regularities.

Similarity: Similarity in analogical processes is traditionally defined by phonological overlap, wherein forms that share segments and/or segmental features are considered more similar (Derwing & Skousen, 1994; Frisch, Pierrehumbert, & Broe, 2004; Frisch & Zawaydeh, 2001; Skousen, 1989, 1993). For the purposes of Arabic, I define similarity at two levels: coarse-grained similarity, which refers to shared structure at the level of the CV template or pattern; and

fine-grained similarity, which refers to shared segmental features beyond those specified by the CV template or pattern. Specifically, the latter is based on the natural class theory developed by Broe (1993) and Frisch, Pierrehumbert and Broe (2004) in which similarity between segments is defined as the ratio of shared to non-shared natural class features. In the computational aspects of this work, the similarity between two forms is calculated using string-edit (Levenshtein) distance (Kruskal, 1983; Levenshtein, 1966) weighted by segmental similarity.

Lexical gang: The lexical gang is a major concept in psycholinguistics, where it is defined as a group of forms with shared structural and segmental features (Albright, 2002; Alegre & Gordon, 1999; Daelemans, Gillis, & Durieaux, 1994; Ernestus & Baayen, 2003). For the purposes of morphology, a lexical gang is one that also displays consistent morphological behavior, similar to Albright's (2002) 'islands of reliability.' The size of a gang is a critical element in generalization, with more examples 'ganging up' to create more support for the analogy (Rumelhart & McClelland, 1986; Stemberger & MacWhinney, 1988). These phenomena are referred to as 'gang effects.' In this work, a gang is defined at the relevant level of morphological abstraction for the system in question. For noun plurals, the primary distinguishing feature among singulars is the CV template, which is supported by the theoretical work in this area (McCarthy & Prince, 1990a). As such, a gang for the noun plural system is defined as a group of singular forms with the same CV template that have shared CV template in plural form (following Dawdy-Hesterberg & Pierrehumbert, 2014). For the masdar system, however, all form I verbs share the same CV template structure. The primary distinguishing feature among verbs is the pattern of the verb. Thus, a gang for this system is defined as a group of verbs with the same pattern that also share the masdar pattern.

1.4 Analogy formation in morphology

Speakers of a language can easily produce morphological variants of a word that they have never before encountered. Since Berko (1958) first demonstrated that both children and adults can successfully form plurals of nonce words, the wug (nonce-form) paradigm has since been widely used in the study of morphological generalization (e.g., Albright, 2002; Berent, Marcus, Shimron, & Gafos, 2002; Bybee & Moder, 1983; Ernestus & Baayen, 2003; Hayes, Zuraw, Siptar, & Londe, 2009; Marcus et al., 1995; Prasada & Pinker, 1993). Generalization to new forms, however, depends greatly on how learnable the existing patterns in the system are, as well as how certain speakers are of what pattern an unseen form should take.

First, what makes a system learnable? From the perspective of the learner, in order to successfully learn a system as a whole, there must be regular correspondences between the input and output forms[1] of the morphological process. That is, if a particular cue (e.g., a phoneme in a particular position in the base form) always results in a particular output pattern, then this pattern should be easily learnable. If, on the other hand, the same cue results in one output pattern 60% of the time and another output pattern 40% of the time, then this system is less learnable, as speakers must learn which output pattern a particular word takes on a case-by-case basis, rather than being able to easily extrapolate from known forms. The predictability of the output based on the cues in the input is critical to learnability.

1. Here I will use the terms 'input' and 'output' to refer to the base and derived/inflected form. I use these terms to highlight the process that the base form undergoes, as well to have one term 'output' that subsumes both derived and inflected forms.

A second factor that makes a system less learnable is the complexity of the changes between the input and output forms. We see some evidence for this, for instance, in the learning trajectories of English-learning children acquiring the past tense of verbs. On the whole, the suffixed plural is learned first, in that it is first to be generalized to new instances. Children at first learn irregular verbs individually and are able to correctly produce them, and then begin to overgeneralize the suffixed plural to irregular verbs once they have learned a sufficient number of suffixed forms to generalize this pattern. Only much later in the developmental trajectory do children re-learn the irregular non-concatenative patterns, such as "fling"⇒"flung" (Berko, 1958; Marcus et al., 1992). Finally, although the time course is not firmly established, speakers can learn to generalize irregular patterns to nonce forms, as in "spling"⇒"splung" (Albright & Hayes, 2003; Bybee & Slobin, 1982; Prasada & Pinker, 1993; Racz, Becker, Hay, & Pierrehumbert, 2014). In this sense, the non-concatenative patterns, which do not have clearly separable past-tense morphemes, are more difficult to learn. This issue is not entirely separable from the relative type frequencies of the regular and irregular plural patterns, however, and so this evidence should be taken with caution. These factors of predictability and non-concatenativity are also at play for children learning Arabic noun plurals, where there is both high unpredictability about the plural for a particular singular and many non-concatenative morphological patterns. As a result, children learning Arabic noun plurals acquire noun plurals much later than children learning concatenative languages (Berman, 1981; Clahsen, Rothweiler, Woest, & Marcus, 1992; c.f. Ravid & Farah, 1999). Further, like English-speaking children, Arabic-speaking children acquire the non-concatenative patterns later than the concatenative ones (Omar, 1973; Ravid & Farah, 1999). The relative type frequencies of the non-concatenative patterns are much higher in Arabic noun plurals, with an estimated 25-40% of noun types taking

a non-concatenative pattern (Boudelaa & Gaskell, 2002; Dawdy-Hesterberg & Pierrehumbert, 2014), than in English past tense verbs, with only 14% of noun types taking a non-concatenative pattern (Marcus et al., 1995). Although the type frequencies of the non-default patterns also play a role in learnability, this is some evidence that non-concatenativity is a source of difficulty in the learnability of morphological systems.

Second, what makes a system generalizable? Once a learner has acquired a system (in the sense that they know the correct output forms for some sufficiently large number of input forms), how do speakers decide which pattern to extent to a new form? Many theories of morphological productivity have argued that generalization to new forms occurs on the basis of analogy to existing forms, such that new forms take the morphological pattern of the most similar existing form(s) (e.g., Prasada & Pinker, 1993; Rumelhart & McClelland, 1986; Skousen, 1989, 1993). Under such a theory, a major issue is in determining how a speaker calculates similarity. Similarity in the domain of phonology is generally defined as the overlap of segmental and prosodic features between two forms (Derwing & Skousen, 1994; Frisch et al., 2004; Frisch & Zawaydeh, 2001; Skousen, 1989, 1993). The exact method of operationalizing similarity in Arabic non-concatenative morphology is a question I take up.

Although analogy as an abstract concept could in principle be to a single form, there is a wealth of evidence that the number of forms taking a pattern has an effect on the likelihood of the analogy (e.g., Albright, 2002; Alegre & Gordon, 1999; Daelemans et al., 1994; Ernestus & Baayen, 2003; Rumelhart & McClelland, 1986; Stemberger & MacWhinney, 1988). That is, the more stored forms taking a particular pattern that are similar to the new form, the more likely the new form is to take that pattern. These effects have been called 'gang effects,' with the idea that

multiple forms 'gang up' to create more support for an analogy than a single form would have. In this sense, larger 'gangs' of forms are more likely to be the basis for analogy.

Thus, under this framework, a speaker needs to determine two things to form a successful analogy for a previously unseen form: first, what is similar to the unseen form; and second, what is likely based on the distribution of existing forms. The selected output pattern for the new form is then a function of these two factors. In order to determine the exact nature of this function, we need to examine two factors in detail. First, as noted above, the basis of similarity in Arabic morphology is not entirely clear. In studies of many concatenative languages, similarity in analogy is largely based on shared segmental and prosodic features. For instance, a nonce form [spling] has significant featural overlap with verbs like [fling] and [cling], and thus speakers are likely to (and have been demonstrated to) form the past tense [splung] on the basis of analogy to similar existing forms (Albright & Hayes, 2003; Bybee & Slobin, 1982; Prasada & Pinker, 1993; Racz et al., 2014). However, in the theoretical literature on Arabic morphology, the structure of the word, namely the CV template and the pattern, are major bases of similarity (McCarthy, 1981, 1993; McCarthy & Prince, 1990a). Moreover, there is computational evidence that this shared structure is the primary basis of analogy in generalizing noun plurals, with additional shared segmental features having only a small influence on analogy formation (Dawdy-Hesterberg & Pierrehumbert, 2014). There is evidence that featural similarity influences word-likeness judgments of nonce verbs in Arabic (Frisch & Zawaydeh, 2001), but there has been no experimental study of the effect of featural similarity versus structural similarity in analogy formation in Arabic morphology. Thus, in non-concatenative morphology, we have two possible measures of similarity: shared structure of the CV template or pattern (depending on the specific morphological system in question), and shared segmental features beyond those specified by the

CV template or pattern. Above, I defined these as coarse-grained and fine-grained similarity, respectively. As shown in Figure 1.3, granularity can be defined as a gradient scale, where more fine-grained means that more segment-level features are specified, and more coarse-grained means fewer features are specified, such that [+cons] is the only feature specified (see also Pierrehumbert, 2001). In Figure 1.3, [s] in "spling" has all of the segmental features specified in the most fine-grained representation on the bottom, while [+cons] is the only feature specified for that segment in the most coarse-grained representation on the top.

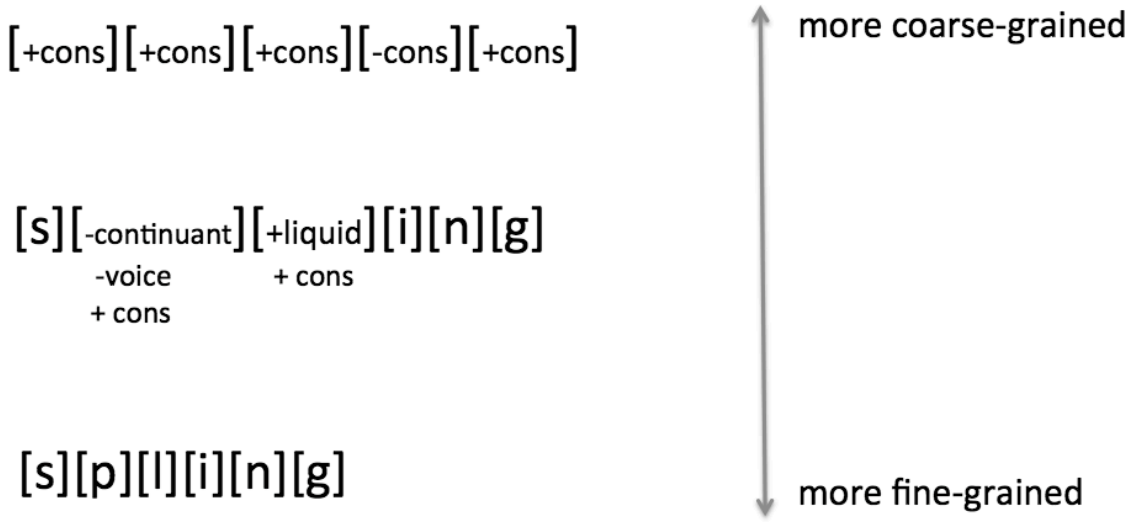

Figure 1.3: Levels of granularity in phonological similarity

As noted, I focus on the coarse-grained level of the CV template and pattern, and the fine-grained level of additional shared segmental features. These are not mutually exclusive; rather, the question is whether segmental features strengthen similarity judgments in analogy formation beyond the similarity defined by the shared word structure of the CV template or pattern. Moreover, it is unclear whether the basis of similarity differs between the two systems under examination, the noun plural and the masdar. For the noun plural, as noted, there is theoretical

and computational evidence that the CV template is the primary basis of analogy formation, and that shared segmental features strengthen an analogy (Dawdy-Hesterberg & Pierrehumbert, 2014; McCarthy & Prince, 1990a). For the masdar of form I verbs, there has been no examination of analogy formation, either computationally or experimentally, and as noted, the CV template is unavailable as a means of distinguishing between verbs as all form I verbs share the same CV template. The potential differences or similarities between these systems will provide interesting insight into the nature of morphological generalization in Arabic as a whole.

The second major question under examination is how speakers select among the available possibilities. The 'optimal' rational strategy would be to select deterministically, or regularize, by always selecting the most-probable option, as this would result in the highest likelihood of accuracy for an unknown form. However, a number of studies have shown that adult speakers often select among the possible choices in a probabilistic manner, producing or selecting a given pattern in proportion to its likelihood (Coleman & Pierrehumbert, 1997; Ernestus & Baayen, 2003; Goldrick & Larson, 2008; Hayes et al., 2009; Hudson Kam & Newport, 2005). This tendency to probability-match depends both on the age of the speaker and the amount of uncertainty in the system, where uncertainty is a function of two main aspects of the system: the number of possible outcomes, and the relative probabilities of those outcomes. The latter dimension has been studied more thoroughly, with many studies varying the relative probabilities of binary outcomes, but a small number of studies have examined systems with three or more outcomes. With regards to age, there is evidence that children tend to regularize when learning an artificial binary-outcome language system with varying probabilities (Hudson Kam & Newport, 2005) as well as when learning an artificial language system with as many as five outcomes (Hudson Kam & Newport, 2009). There is also limited evidence from natural

language acquisition that children tend toward regularization when exposed to inconsistent input (Singleton & Newport, 2004). For adults, the tendency toward probabilistic versus deterministic behavior is modulated by the amount of uncertainty in the system. In artificial language studies, adults tend more toward regularization when there is a larger number of possible outcomes (while holding the proportion of the primary outcome constant) (Hudson Kam & Newport, 2009) as well as when there is both variability in the input as well some bias toward a particular outcome, presumably stemming from the L1 (Schumacher, Pierrehumbert, & LaShell, 2014; Wonnacott, Newport, & Tanenhaus, 2008). In addition, there is some evidence in both the psychological and the linguistic domains that adult individuals have different tendencies toward probabilistic or deterministic behavior in category learning and generalization that is independent of the degree of consistency in the input (Hudson Kam & Newport, 2005; Nosofsky & Johansen, 2000; Schumacher et al., 2014; Wonnacott & Newport, 2005). While Nosofsky and Johansen theorize that these behavioral differences stem from individuals placing different attentional weights on the varying dimensions of the input stimuli, the underlying mechanism for these individual differences remains unexplained. Nonetheless, this observation that individual differences may also influence generalization is relevant to the task at hand, and will also be examined in this thesis.

However, there has been little if any examination of speakers behavior in natural language tasks in which there are a large number of possible morphological variants, as the majority of these studies have used either artificial language paradigms that manipulate the amount of inconsistency, or natural-language systems with a binary choice. Thus, Arabic provides an ideal natural language test case for studying how speakers select among possible

morphological variants when there are a large number of possible outcomes and high uncertainty about the optimal choice.

1.5 Roadmap

In this thesis, I will examine learnability and generalization of the morphology of Modern Standard Arabic, focusing on the noun plural and masdar systems. I will use psycholinguistic experiments and computational analyses to assess two major aspects of generalization. First, I will address the learnability of a morphological system based on the predictability of the morphological variant of an unseen form based on analogy to existing forms that are available in a speaker's lexicon. In doing so, I will address what types of linguistic information are available in the input as cues to the output for a particular word undergoing some morphological process.

When speakers form new words based on analogy to existing words, how do they draw this analogy? As noted, there is evidence that the basis of similarity in Arabic is different than in languages with concatenative morphology, with the primary basis of analogy in Arabic being shared structure (namely the CV template for the noun plural system), with a small influence of shared segmental features (Dawdy-Hesterberg & Pierrehumbert, 2014; McCarthy & Prince, 1990a). In contrast, in languages with concatenative morphology like English and Dutch, shared segmental features play a relatively larger role in analogy formation (e.g., Alegre & Gordon, 1999; Ernestus & Baayen, 2003). Thus, a major question under investigation is whether the basis of analogy differs between the noun plural and the masdar systems, and how these differences or similarities reflect aspects of morphological generalization in non-concatenative morphology in general.

Second, I will assess how speakers generalize existing morphological patterns to previously unseen forms using nonce-form tasks. By presenting speakers with non-existing but Arabic-like forms and asking them to create a noun plural or masdar, we can gain insight into how speakers determine the best morphological pattern for an unseen form. By comparing these experimental results to computational models of analogy with varying bases of similarity, we can find converging evidence on how speakers determine similarity in creating analogies for unseen forms.

More generally, this thesis investigates how speakers generalize morphological patterns in systems with two main characteristics: coarse-grained representations, as both systems contain a number of non-concatenative patterns that require a high level of abstraction to represent and generalize; and high uncertainty, in that there are 30+ patterns for both of the systems under investigation. By studying systems with both of these characteristics, I will examine the key questions of: 1) what is the basis of analogy in morphological generalization in Arabic?; and 2) how do speakers decide among the possible outcomes when there are a large number of possibilities? These questions speak both to issues in Arabic linguistics and to psycholinguistics more generally.

In chapter 2, I will examine the noun plural system. The noun plural is relatively well-studied, and the basis of plural formation is generally understood. However, there has been little, if any, study of how native speakers generalize these patterns to previously unseen forms. Although it may intuitively seem to be the case, the models that best predict a linguistic system do not always predict speaker behavior well (e.g., Becker, Ketrez, & Nevins, 2011; Gagliardi, Feldman, & Lidz, 2012; Gagliardi & Lidz, 2014), as speakers may under- or over-rely on linguistic cues that the model takes to be equal. In comparing the results of native speaker

pluralizations of nonce forms to the plurals predicted by the most-accurate models of the noun plural system, I will examine which linguistic cues speakers can, and do, attend to in learning, and then generalizing existing plural patterns.

In chapter 3, I will examine the masdar (verbal noun) system. The masdar system of form I verbs is understudied, and in fact has been frequently cited as unpredictable (Grenat, 1996; Holes, 2004; Kremers, 2012; McCarthy, 1985; Ryding, 2006). In this thesis, I will use analogical modeling on a set of existing verb-masdar pairs to show that there are regular phonological correspondences between the verb and the resulting masdar form.

In chapter 4, I will use two psycholinguistic experiments to examine speaker knowledge of the masdar system. The first experiment examines the issue of verbs that have multiple masdars using a forced-choice experiment asking speakers to select the preferred masdar for these verbs. Although dictionary sources claim that these verbs have multiple active masdars, it is difficult to discern whether both of these forms are truly active in the language, and if so, why multiple forms are available. In a second psycholinguistic experiment, I will examine how native speakers generalize existing masdar patterns to nonce verbs. As with the experiments on the noun plurals, this will give insight into whether the available linguistic cues (based on the modeling work in chapter 3) are truly utilized by speakers in creating new masdars for previously unseen verbs.

Finally, in chapter 5, I will discuss the similarities and differences between the two morphological systems, and how the results of the computational and experimental work give insight into issues in Arabic linguistics, and to more general issues in learnability and generalization.

Chapter 2 : Generalization of Arabic noun plurals

2.1 Introduction

The Arabic noun plural system provides an excellent forum for examining keys issues in morphological generalization for a number of reasons. First, previous research on the noun plural system has found that the linguistic factors available to speakers in the lexicon are only partially determinate of the plural. Specifically, analogical modeling work has found that the plural can be predicted accurately for unseen forms based on existing forms with only 65-70% accuracy, where a random type-frequency-weighted baseline achieves 39% (Dawdy-Hesterberg & Pierrehumbert, 2014; Nakisa et al., 2001; Plunkett & Nakisa, 1997). One identifiable issue is that the primary plural pattern, the [-aat] suffix, is not overwhelmingly in the majority, with 59-74% of noun types taking this pattern in corpus analyses (Boudelaa & Gaskell, 2002; Dawdy-Hesterberg & Pierrehumbert, 2014). Second, there are a large number of possible plurals in the system, with some scholars identifying as many as 33 distinct plural patterns (Levy, 1971; McCarthy & Prince, 1990a; Wright, 1988). Both of these factors lead to a great deal of uncertainty on the part of the speaker in choosing the best plural for an unseen form.

The literature has identified a few major factors that partially predict the plural form of a given noun singular. The CV template of the singular has been shown to be the primary determinant of the plural. First, only a subset of singular templates take broken plurals, which are termed 'canonical' by McCarthy and Prince[2] (1990a). Singular templates that do not have canonical structure virtually always take sound plurals (Dawdy-Hesterberg & Pierrehumbert,

2. Note that canonicity is not defined by whether a singular template can take a broken plural, but rather by the prosodic minimal word in Arabic as defined in McCarthy and Prince (1990a). The singular templates examined in this series of experiments all have canonical structure and thus are eligible for both broken and sound pluralization.

2014; McCarthy & Prince, 1990a), which means that plurals for these singulars are extremely predictable. For singulars with canonical structure, only a subset of the broken plurals are attested for a given singular template, thus limiting the number of patterns from which to select a plural (Levy, 1971; McCarthy & Prince, 1990a; Ratcliffe, 1998). In addition, the CV template of the plural is partially determined by the CV template of the singular, so plural CV templates that do not match the singular template in the relevant features should not be eligible candidate templates. For example, in quadrilateral plurals, the moraic weight of the final syllable is maintained in pluralization, as in [jun**dub**]⇒[janaa**dib**] ("grasshopper") vs. [sul**tˤaan**]⇒[salaa**tˤiin**] ("sultan").

From studies of loanword assimilation, we can observe the influence of the CV template on noun pluralization in action. Although many loanwords take the majority sound [-aat] plural, given canonical structure, loanwords also take a variety of broken plurals, for example, in [film]⇒[ʔaflaam] ("film"), [bilyuun]⇒[balaayiin] ("billion"), [sˤandal]⇒[sˤanaadil] ("sandal"), and [muwtuur]⇒[mawaatiir] ("motor").

In addition to the CV template of the singular, the other major factor in pluralization is the distribution of existing noun types taking a given plural template (Dawdy-Hesterberg & Pierrehumbert, 2014). A singular noun is most likely to take the most-frequent plural template among forms that have the same singular template structure. That is, type statistics on nouns with the same singular template, not across all singular nouns, are the major predictor of the plural for a given singular.

Finally, analogical modeling has also shown that there is phonological regularity in the system beyond that defined by the CV template. A singular is more likely to take a particular plural if it also shares segmental features with existing singulars taking that plural (Dawdy-

Hesterberg & Pierrehumbert, 2014). Although Dawdy-Hesterberg and Pierrehumbert did not examine which segmental features were most predictive of the plural, this finding corroborates some of the theoretical literature that notes some segmental features being relevant to plural formation. Specifically, the voweling of the singular and the presence of 'weak' verbal root consonants ([j], [w], and [ʔ]) have been posited to play a role in plural formation (Ratcliffe, 1998).

Two other relevant factors in pluralization are the animacy and gender of the word, which in tandem mediate the choice between the two sound plurals. 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. The semantic features governing this alternation are well understood, and so this analysis focuses on the morphophonological and distributional properties of the system. Henceforth, "sound plural" will refer to the [-aat] plural unless explicitly noted.

The studies presented in this chapter examine native speaker pluralization of nonce singular nouns. Using a nonce-form (wug) paradigm, these experiments ask whether speakers use the same linguistic features identified in the modeling work in determining the best analogy to form a new plural from an unseen form. The experiments use nonce singulars with eight different template structures, all of which are strongly attested in the Corpus of Contemporary Arabic (Al-Sulaiti, 2009). These were selected from among the gangs in Dawdy-Hesterberg and Pierrehumbert (2014). For four of the singular templates, the majority of forms in the dataset take a broken plural, while for the other four, the majority of forms take a sound [-aat] plural. Of the templates taking a majority broken plural, each takes a different plural template. The first experiment, 1A, uses an open-response paradigm, in which speakers respond with whatever

comes to mind as the plural for a given nonce form. The second experiment uses a forced-choice paradigm, in which speakers select from two possible plurals. In experiment 1B, the plural options are collected from the plurals given by participants in the open-response paradigm. Using the open-response paradigm, we can ascertain first how speakers form plurals when all options are theoretically available to them. Then, using the forced-choice paradigm, we can examine how speakers select among possible plurals when the options are constrained. Using both the open response and the forced-choice paradigm allows us to examine whether there are task differences in selecting among morphological variants when the options are constrained versus unconstrained.

For experiment 1A, I rely on a comparison of analogical models to behavioral data in order to best determine the linguistic factors that speakers are attending to in pluralizing nonce forms in Arabic. First, I use an open-response nonce-form task to elicit plurals for unseen singular nouns. Second, I compare the plurals given by participants to the plurals predicted by four analogical models. For experiment 1B, in which the task is a forced choice among the possible plurals from experiment 1A, I will examine the extent to which speaker preference for the nonce plurals is predicted by probability of the plural from corpus estimates and the probability of the plural as a response in experiment 1A, and possible interactions between these probabilities. Finally, I will discuss the implications of the results and future directions.

2.2 Experiment 1A

2.2.1 Methodology

2.2.1.1 Participants

Participants were recruited via Amazon's Mechanical Turk, a marketplace for online tasks that has been successfully used for psycholinguistic research in recent literature (Crump, McDonnell, & Gureckis, 2013; Schnoebelen & Kuperman, 2010; Sprouse, 2011). The heading for the experiment was "Answer a survey about Arabic words" (in Arabic). Participants received $5 upon completion of the experiment. In total, 904 participants accepted the task on Mechanical Turk, of which only 171 completed any experimental items. 87 participants completed the entire experiment. 26 of those participants were excluded from analysis for the following reasons: non-native speaker of Arabic (n=3), achieving less than 80% accuracy on filler items (n=17), database error which resulted in not seeing all items (n=4), for having completed experiment 1b previously (n=1), or for having a large proportion of the nonce items excluded (n=1). In total, 61 participants completed the experiment and met all qualifications for inclusion in analyses.

Of the 61 participants whose data was analyzed, 36 participants were male and 19 were female. Gender was not recorded for 6 participants due to a database error; all other demographic information was recorded for those participants so they were included in analyses. All participants were self-reported native speakers of Arabic. Mean proficiency in MSA was 8.97 on a scale of 1-10 (S.D.=1.68). Mean frequency of use of MSA was 6.89 on a scale of 1-10 (S.D.=2.38), with 1 being "rarely use MSA" and 10 being "use MSA frequently." For level of education, 7 participants reported having less than a college education, 40 participants an undergraduate education, 12 participants a master's degree, and 2 participants a doctorate. All but one participant reported also speaking English as a second language, with a mean proficiency of 8.10 (S.D.=1.77) on a scale of 1-10. 27 participants reported speaking a third language, and 11 reported speaking a fourth.

Information on primary spoken dialect was also elicited. Dialects were classified by major regional dialect. 26 participants reported speaking Egyptian as their primary dialect. 12 participants reported speaking a Levantine dialect (includes Jordanian, Syrian & Lebanese). 11 participants reported speaking a North African dialect (includes Moroccan & Tunisian). 7 participants reported speaking a Peninsular dialect (includes Bahraini, Emirati, Omani & Yemeni). 2 participants reported speaking a Mesopotamian dialect (both Iraqi), 1 participant reported speaking Sudanese, and 2 participants did not specify a dialect. All participants were analyzed in the main results, but only dialect groups including at least 10 speakers were used in the dialect analysis.

2.2.1.2 Experimental materials

2.2.1.2.1 Stimulus design

This experiment examines eight singular templates that are strongly attested in the corpus set. All eight singular templates have canonical structure, and thus are eligible for broken pluralization. Four of these singular templates take a broken plural as the majority pattern, and four take the sound [-aat] plural as the majority pattern. All of the singular templates have at least two attested plural templates, with some having as many as eight. All comparison statistics are from the set of 1945 singular-plural pairs used in Dawdy-Hesterberg & Pierrehumbert (2014). The sound plurals in this set were collected from the CCA, and the broken plurals were provided by Mohammed (2009) and cross-checked against the CCA. All pairs were hand-checked using the Buckwalter Arabic Morphological Analyzer (Buckwalter, 2004; Dehdari, 2009).

48 filler items were selected from this set, and were all moderate-to-high frequency existing singular nouns that take the dominant plural template for that singular template. Filler

items were also used as qualifying questions to ensure that participants were proficient in Arabic and completed the task as instructed. For each filler item, there were five matched nonce items, for a total of 240 nonce items. Each participant saw one of five sets, with the sets counterbalanced across participants.

The nonce items were constructed by creating all possible permutations of non-existing triliteral verbal roots using the set of existing triliteral roots from Pierrehumbert (1993). These roots were then filtered for OCP-Place violations (Frisch et al., 2004; Frisch & Zawaydeh, 2001; McCarthy, 1986). Phonotactic probabilities were computed for each nonce root based on probability of occurrence of each root consonant in that position in existing roots, such that *p(root)* = *p*(C1)**p*(C2)**p*(C3) (Frisch & Zawaydeh, 2001). Neighborhood densities were also calculated, where a neighbor is defined as an existing root sharing two consonants in any position (Frisch & Zawaydeh, 2001). The nonce roots were matched to the filler roots for phonotactic probability and neighborhood density, and nonce and filler roots were not significantly different in these measures. The selected roots were also cross-checked against the Aralex database (Boudelaa & Marslen-Wilson, 2010) and by a native Arabic speaker to ensure that they did not occur in any existing forms. Finally, the nonce roots were inserted in the pattern of the corresponding filler item. For instance, for the filler item [maktab], the pattern was [ma$C_1C_2aC_3$]. One matched nonce root was [ʃmk], resulting in the nonce singular [maʃmak]. Roots were also matched for 'weak' root consonants ([w], [j], and [ʔ]); for example, all roots for the singular pattern [CvvC] had a weak consonant in the C_2 position.

2.2.1.2.2 Procedure

Participants first saw an introduction screen with a short explanation of the experiment and an example. If participants accepted the task on Mechanical Turk, they then viewed the consent form and clicked "decline" or "consent." Participants who gave consent continued to a page of questions on demographic information and language background. Following this, participants saw instructions with two examples, followed by a short practice section with 4 sample items. Finally, participants entered the test section.

The test section consisted of 96 items, with 48 filler items and 48 nonce items. Participants who did not correctly pluralize 80% of filler items were excluded. Analyzed participants had a mean accuracy of 87.2% on filler items (S.D.=4.5).

Items were presented one at a time in two-sentence frames. The singular form always occurred in the first sentence, and was marked in blue (grey in the example stimulus below). The second sentence contained a blank that syntactically required a plural. Below the sentences, there was a text input box where participants typed in what they would put in the blank. Participants could not continue to the next page until they entered text with at least one short vowel diacritic. Figure 2.1 shows a sample stimulus.

Two sentence frames were created for each filler item, and nonce items were presented in the sentence frame of the filler item to which they were matched. Sentence frame was counterbalanced across participants. Order of presentation was randomized for each participant.

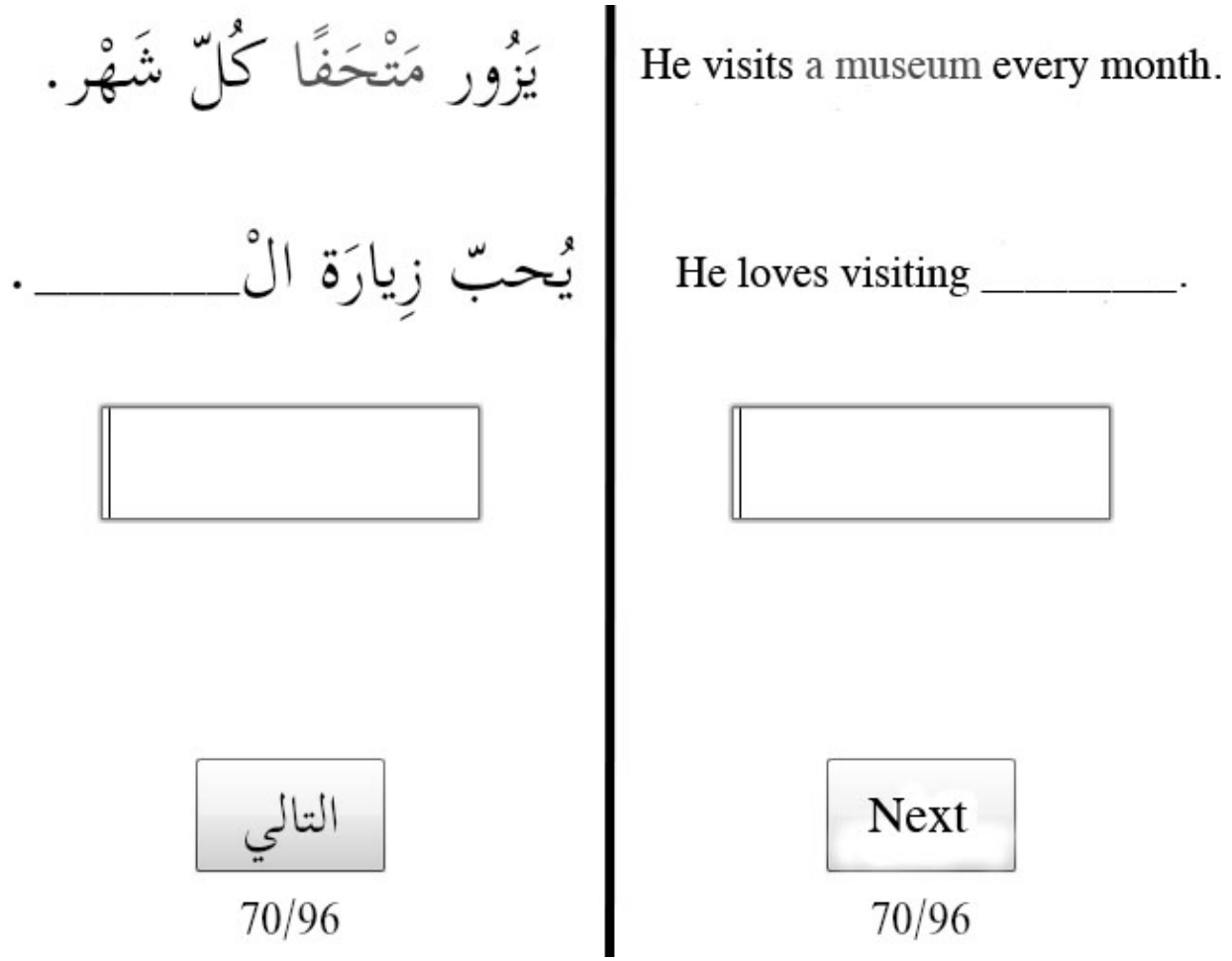

Figure 2.1: Example filler item from experiment 1A (left) and English gloss (right)

All experimental materials were in written Modern Standard Arabic. Dialect-specific variants were avoided. All practice and test items were presented in fully diacritized form, with all short vowels and geminates marked on target singulars for phonological transparency, and on all forms in frame sentences as necessary for comprehension. The consent form and instructions were not diacritized, aside from example items. Participants were instructed to input the plural in fully diacritized form.

2.2.1.2.3 Response Coding

Participant responses to test items were coded for plural type (broken or sound), prosodic structure (McCarthy & Prince, 1990a), CV template, pattern, whether the plural template occurred in the set of existing plurals (expected vs. unexpected), and the ranking of the plural template in the set of existing forms. Filler items were coded for accuracy only.

First, responses were stemmed for: the definite article [al], possessive pronoun suffixes, and case markings. The responses were then converted to pattern, such that consonants occurring in the singular were replaced by C but vowels and additional consonants in the plural pattern were maintained (ex: [ʔahliqaT] ⇒ [ʔaCCiCaT] from singular [halq]). These were checked by hand to ensure that the patterns had licit syllable structure.

One issue in coding is that there is significant variability in diacritization of written Arabic (Buckwalter, 1997). Even when explicitly instructed to do so, many participants did not include a diacritic on every base character. However, full diacritics are often not necessary for disambiguation given restrictions on syllable structure in Arabic. Thus, the following procedure was followed in coding. If a base consonant did not have a diacritic, it was assumed that it was unvoweled if it was a consonant, and that it was a long vowel if it was [w], [j] or [ʔ]/[A]. Additionally, [h] in final position was considered to be [T] if and only if there was a short [a] on the previous character and [h] was not in final position in the singular. Finally, if [ʔ]/[A] was in initial position but unvoweled, it was considered to be [ʔa], as this is the most common voweling for this character, and there are a number of common plural patterns that begin with this sequence (e.g., [ʔaCCiCaT], [ʔaCCaaC]).

The pattern for each response was converted to CV template by replacing short vowels with [v] and long vowels with [vv]. The broken plurals were coded by prosodic pattern, where

iambic is [CvCvv+], trochaic is [CvCvC] (plus optional suffixes), monosyllabic is [CvCC] (plus optional suffixes), and 'other' is [CvCCvvC] and [CvCCvC] (McCarthy & Prince, 1990a). Any broken plural responses that had licit syllable structure but did not conform to one of these categories was marked 'unknown.' Finally, each response was marked according to whether it occurred in the corpus set from Dawdy-Hesterberg & Pierrehumbert (2014) (expected vs. unexpected), and the ranking of the template in the corpus set (e.g., sound for the sound plural, broken1 for the most frequent broken plural, broken2 for the second most frequent, etc.).

Responses were excluded from analysis for the following reasons: identity to singular, metathesis of singular root consonants (ex: [jaxaam]⇒[xajaamaat], substitution of singular root consonants (excluding those additional pattern consonants in attested plural patterns), double plural marking (broken stem change + sound plural suffix), and illicit syllable structure. In total, 10.9% of responses were excluded from analysis (n=288).

2.2.2 Results

2.2.2.1 Overall results by singular template

The overall results for participant plural CV templates by singular template are shown in figure 2.2. Each bar represents the distribution of plural templates for that singular template, with sound plurals in white and broken plurals in gray. Broken plurals are divided by CV template. For all templates, we find that broken plurals are used the majority of the time for the four leftmost singular templates, which have a broken plural as the dominant plural in the lexicon. This pattern is reversed for the four rightmost singular templates, which have a dominant sound plural in the lexicon. Further, we find that the dominant plural CV template for a given singular

template is used more often than any other plural template, with 61% of forms on average taking the dominant CV template. This is quite variable across singular templates, with as few as 41% of [CvvC] forms and as many as 86% of [CvCvvCvT] forms taking the dominant plural template.

Responses overall are far from categorical, and some singular templates were given as many as 21 different plural templates. For singular templates that are trilateral (having three consonants), we find a wider variety of broken plurals than for the quadriliteral singular templates (those having four consonants; [CvCCvC] and [CvCCvvC]). This is not very surprising, as the set of possible plural templates for quadrilateral plurals is much more restricted. However, there are instances in which a participant used a trilateral plural for a quadrilateral singular, and vice versa, by omitting or repeating consonants in the singular, for example, [xajaajir] from singular [xajr], when [CvCvvCvC] is not an attested plural for the singular template [CvCC].

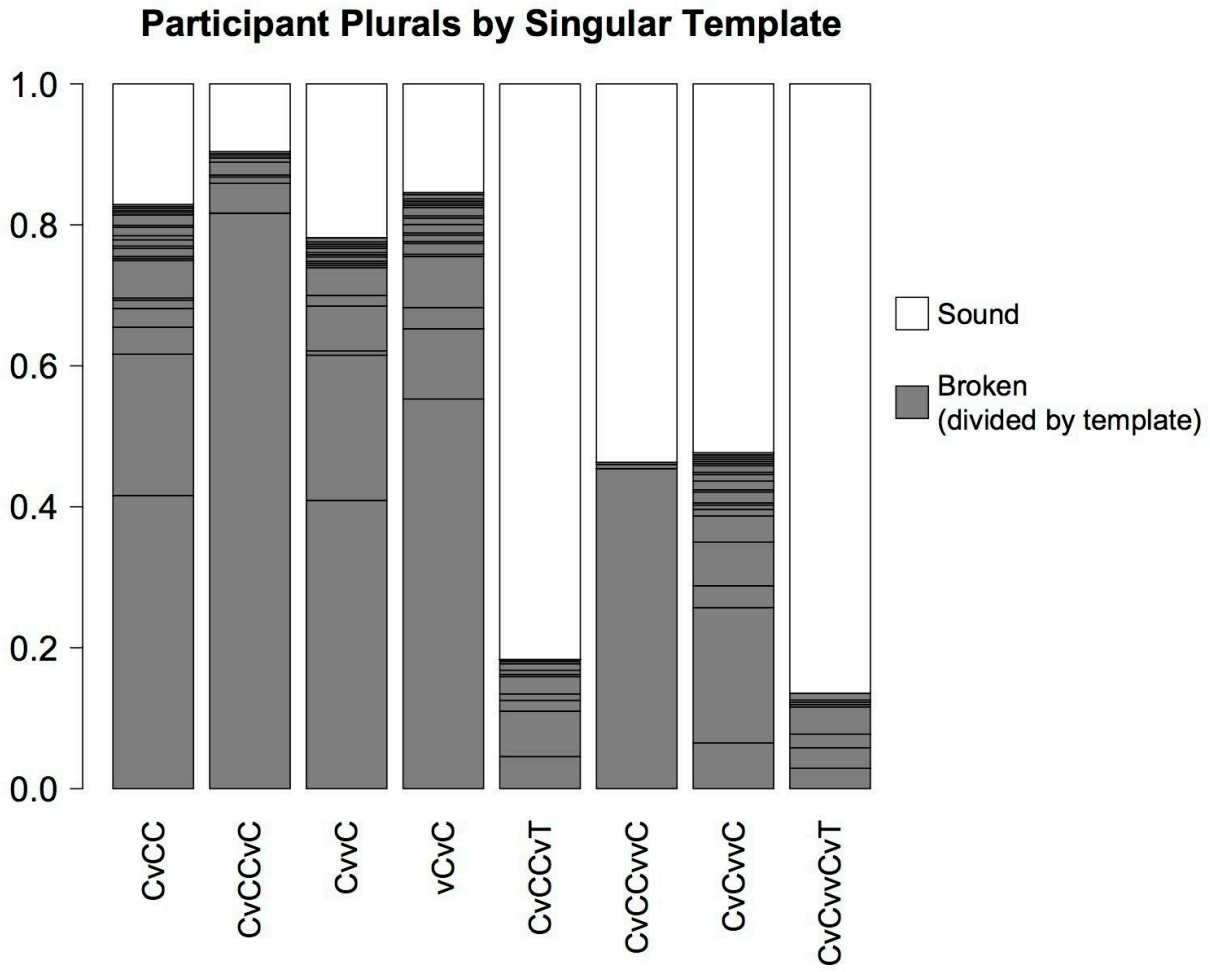

Figure 2.2: Plural CV templates by singular template

Using the set of 1945 singular-plural pairs from Dawdy-Hesterberg & Pierrehumbert (2014) expected probabilities for each plural template by singular template were computed, such that the expected probability is the proportion of forms with a given singular template taking a given plural template. The observed probabilities were calculated in the same manner from the set of participant responses. For any plural that does not occur in the corpus set but is given in participant responses, Laplace smoothing was used (Lidstone, 1920), such that a small but non-zero probability of 0.001 is assigned. Figure 2.3 shows the expected versus observed log

probabilities and the best-fit line. The expected probability is significantly positively correlated with observed probability, $r=0.663$, $p<0.001$. If only plurals that appear in the corpus with that singular template are considered (predicted plurals), the expected probability is nearly identical, $r=0.662$, $p<0.001$. Plurals that were not predicted based on the corpus set are very infrequent in the participant responses, and the observed probabilities for the plurals that were predicted track fairly closely with expected probabilities.

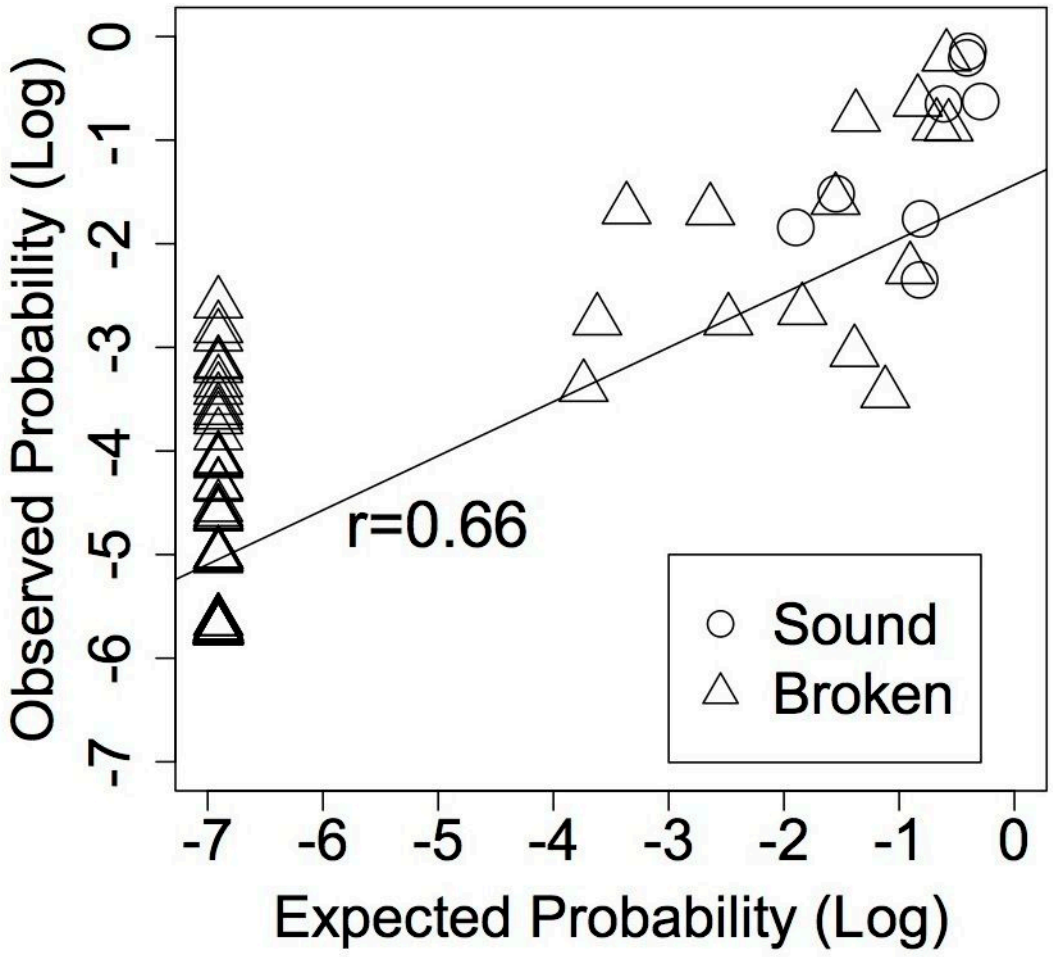

Figure 2.3: Expected vs. observed log probabilities for plural CV template by singular CV template

The aggregate data across singular templates does not reflect how little agreement there is across participants for individual nonce forms. Out of 240 nonce items, only 18 show 100% agreement on the plural CV template across participants. For some items, responses include as many as seven plural templates. For example, the nonce singular [halq] generated responses with six broken plural templates as well as the sound plural: [ʔahliqaT], [ʔahalaaq], [ʔahaaliyq], [hilqaan], [halaaʔiq], [huluwq], and [halqaat]. Overall, 60% of items were given 3 or more plural templates.

In an analogical framework, the plural is selected on the basis of similarity to existing items. Although each group of singulars has a common CV template, the singulars differ in segmental features beyond those defined by the CV template. The items were controlled for weak consonants ([w], [j], and [ʔ]) and for phonotactic probability and neighborhood density, but there is still a great deal of variation in the vowel and consonant qualities within each group of singulars. Thus, some of the disagreement across items within templates may stem from the differing segmental similarity to existing items. However, this does not explain the disagreement within items across speakers. One possibility is that the disagreement stems from differences between participants in dialect background, as the Arabic dialects are fairly divergent in some properties. This is explored next.

2.2.2.2 Analysis of dialect background

Noun pluralization is relatively consistent across Arabic dialects, with broken and sound plurals occurring in virtually every major dialect (Holes, 2004). However, Arabic dialects vary greatly in a number of dimensions related to morphology, most notably in phonological

inventories and phonotactics. The observed phonological differences between MSA and individual dialects in some cases lead to different template structures in MSA and the dialect. For example, complex onsets are disallowed in MSA, but are licit in Moroccan Arabic (Harrell, 1962). In Moroccan Arabic, short vowels in initial syllables are commonly deleted, as in MSA [kitaab] 'book' versus Moroccan [ktaab]**.** Thus**,** it is possible that differences in pluralization might arise due to these phonological and phonotactic differences.

Participants were classified by major dialect region. The dialect regions examined here are: Egyptian (n=24), North African (n=11) and Levantine (n=11). The distribution of plurals for each regional dialect is shown in Figure 2.4. There are some differences between dialects, but aggregate results across participants hide the inter-speaker variation within each dialect. To assess whether dialect background had a consistent effect on pluralization, Krippendorff's alpha (Krippendorff, 1980) was used to measure inter-speaker agreement within each dialect group and across all dialect groups. This coefficient computes the overall agreement above chance between raters on assigning *n* items to *c* categories, where 0 = no agreement at all and 1 = perfect agreement. In the case of this experiment, a 'rater' is a participant, an 'item' is a nonce singular, and a 'category' is an output plural template. Thus, this coefficient represents a measure of agreement between participants on plural CV templates for the nonce singulars.

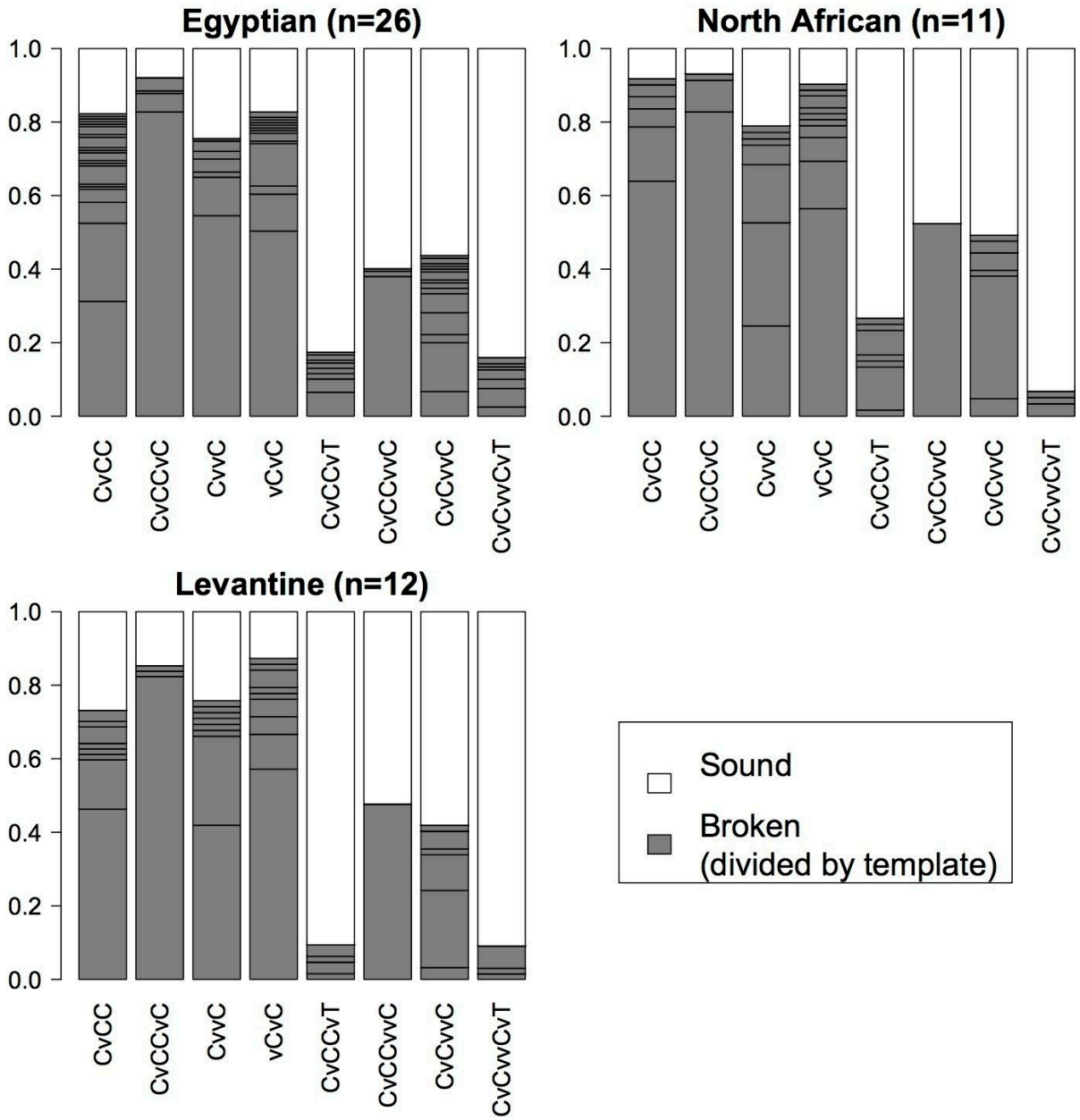

Figure 2.4: Distribution of plurals by singular template across regional dialects

Across all dialect groups, α=0.368. This confirms that, as noted above, participants in general show very low agreement on the CV template for a given singular. Within each dialect group, the agreement is also very low: Egyptian, α=0.356; North African, α=0.464; Levantine, α=0.404. In total, the agreement within dialect group is very similar to the agreement across

dialects. The agreement for the North African group does appear to be slightly better than for the other groups; however, the overall agreement is nonetheless very low, so this was not investigated further. Because the differences between dialects do not seem to have a consistent effect on pluralization, all further analyses are collapsed across dialects.

2.2.3 Comparison to models of pluralization

2.2.2.3.1 Model details

Although I demonstrated above that plural frequencies in the lexicon are strongly correlated with plural frequencies in the experimental data, this does not account for all of the variability observed in the data. In order to determine which factors best account for the variability, four predictive analogical models are compared to the experimental data to determine how speakers select the plural for a new form.

There are two factors in analogy formation examined here. First, what are the morphophonological features used by speakers in defining similarity to existing forms? This questions asks whether speakers use coarse-grained similarity on the level of the CV template, where all singular forms with that singular template are considered equally similar, or use additional fine-grained similarity, such that forms with the same singular template that also contain more similar segments are weighted more heavily in analogy formation. Second, given a set of possible plurals, what decision rule do speakers use? This question asks whether speakers choose deterministically or probabilistically among candidate plural templates. Below, I outline four predictive analogical models that examine these features in a 2x2 design. I then compare the plural templates selected by each model to the plural templates in participant responses to

determine which factors participants are using in generalizing existing plurals to the nonce forms.

All of the models presented here use an analogical framework, in which a test item (the nonce singular) is compared to a set of existing singulars (the comparison set), and the test item is predicted to take the plural pattern of the most similar gang from the comparison set. As a reminder, a gang is a set of forms that have consistent morphological behavior. As stated in the introduction, I define a morphological gang for the noun plural as a set of forms with the same singular CV template that also take the same plural CV template, following Dawdy-Hesterberg & Pierrehumbert (2014). The similarity of a test singular to a gang is weighted by the number of forms in the gang, as type frequency has been shown to have an important role in morphological analogy (Albright, 2009; Baayen & Lieber, 1991; Bybee, 1995). The definition of similarity is varied amongst the models as noted above, where in two of the models similarity is defined by a shared CV template, and in the other two models similarity is defined using the CV template plus additional fine-grained segmental features, such that forms with similar segments beyond the CV template will also be considered more similar by the model. The decision rule used by the models is also varied. The specific parameters of the models and methods of implementation are discussed below.

One shortcoming of these models is that the existing words over which they form analogies were drawn from an undiacritized corpus (Al-Sulaiti, 2009). In undiacritized (so-called "unpointed" or "unvoweled") orthography, the diacritics that mark short vowels are omitted, as well as the diacritics marking geminates. Thus, the nonce forms were also presented to the models in undiacritized form, so that the orthography between the comparison and test items was consistent. This means that the models had access to the quality of the long vowels and the

consonants in the singular forms, but not the short vowel quality. The voweling of the singular has been reported to be a secondary factor in noun pluralization (Levy, 1971; Ratcliffe, 1998), but the omission of the short vowels may have an impact on the role of segmental similarity in analogy formation. The results of the model comparison will be discussed in light of the potential shortcomings of using undiacritized text.

Simple Template Match: This model uses type statistics on the singular CV template to determine the best gang with which to form an analogy, and selects from the candidate gangs deterministically (Dawdy-Hesterberg & Pierrehumbert, 2014). Thus, the model always selects the largest gang with a same singular template as the test item. Because this model is deterministic, the same gang will always be selected for a given singular, so the predicted plurals are generated from a single iteration of the model.

Probabilistic Template Match: This model also uses type statistics on the singular CV template, but uses a probabilistic choice rule. The model considers only gangs with the same singular CV template as the test item, and outputs a probability for that gang based on the number of items in the gang relative to the number of items in all gangs with the same CV template. That is, if there are two gangs with the same singular template as the test item, with 10 forms in gang A and 5 forms in gang B, the model will assign gang A a probability of 66.7% and gang B a probability of 33.3%. This model thus generates a probability distribution over plurals for each nonce form.

Generalized Context Model, Template-Restricted: This model is a variation of the Generalized Context Model (Nosofsky, 1990) as used in Albright and Hayes (2003) and Nakisa, Plunkett and Hahn (2001). This model has been adapted for Arabic morphology by adding an additional parameter that restricts the candidate set of gangs to only those that have the same singular template as the test item (Dawdy-Hesterberg & Pierrehumbert, 2014). Similarity of the test form to each form in each candidate gang is calculated using string-edit (Levenshtein) distance (Kruskal, 1983; Levenshtein, 1966) that incorporates gradient segmental similarity based on shared natural classes (Broe, 1993; Frisch et al., 2004). String-edit distance is transformed to shared similarity using the transformation in equation 2.1, where *dij* is the distance from test form *i* to comparison form *j*, and *s* and *p* are fitted parameters set to 0.3 and 1, respectively, following Dawdy-Hesterberg & Pierrehumbert (2014) and Albright and Hayes (2003).

Equation 2.1: String-edit distance to similarity transformation

$$n_{ij} = \exp\left(\frac{-d_{ij}}{s}\right)^{p}$$

Similarity for each gang is the summed similarity of each member of the gang to the test form, divided by summed similarity of all forms in all gangs to the test form. This model chooses deterministically, so it always chooses the gang with the highest similarity score to the test item.

Probabilistic Generalized Context Model, Template-Restricted: This model calculates similarity between a test item and a gang in the same manner as the *GCM, Template-restricted*

described above, but uses a probabilistic choice rule rather than a deterministic one. As with the other three models, the candidate gangs are those gangs that have the same singular CV template as the test item. The probability for a gang is calculated as the similarity measure for that gang, divided by the summed similarity measures of all gangs in the candidate set. For example, if gang A has an similarity measure of 4.5 and gang B has a similarity measure of 2, gang A will be assigned a probability of 4.5/4.5+2 = 69.2%, while gang B will be assigned a probability of 2/4.5+2 = 30.8%. This model generates a probability distribution over plurals for each nonce form.

2.2.2.3.2 Model fitting procedure

The distributions of gangs predicted by the four models described above were compared to the distributions of plurals in participant responses using the Jensen-Shannon (J-S) divergence (Cover & Thomas, 1991), which is a measure of the difference between two probability distributions. The J-S divergence is a symmetric, weighted average of the Kullback-Leibler (K-L) divergence (Kullback & Leibler, 1951), where the K-L divergence is the expected number of additional bits required to encode probability distribution *Q* using distribution *P*. Using the base 2 logarithm, $0<D_{JS}<1$. The J-S divergence between *P* and *Q* is given in equation 2.2.

Equation 2.2: J-S divergence between distribution *P* and distribution *Q*

$$D_{JS}(P||Q) = \frac{1}{2}D_{KL}(P||M) + \frac{1}{2}D_{KL}(Q||M)$$

where

$$D_{KL}(P||Q) = \sum_i \log\left(\frac{P(i)}{Q(i)}\right) P(i)$$

and

$$M = \frac{1}{2}(P + Q)$$

The distributions from all four models were first corrected using Laplace smoothing (Lidstone, 1920), such that any gang that was not predicted by the model was assigned a small but non-zero probability of 0.001. The divergence was then calculated between each model and the experimental results for each item across participants, and for each participant across items with the same singular template. The divergences were then averaged for each model for each singular template, across items in the by-item comparison and across participants in the by-participant comparison. These divergences by singular template were then averaged for each model, for an aggregate value of how well each model fit the data overall. The best-fitting model is the one with the lowest divergence from the experimental data.

2.2.2.3.3 By-item fit

First, I will examine the model fit to the participant data by item. Table 2.1 below shows the mean divergence between each model and the experimental data by item. Probability

distributions on the plural CV template were calculated for each item across participants, and for each model for each item. There were 240 nonce items, resulting in 960 comparisons between the experimental data and the models. The divergence was averaged across items by singular template (shown in rows labeled by singular template), and then averaged for each model across singular templates (shown in row labeled "mean"). As noted, with the J-S divergence, the lower the divergence, the better the fit between the two distributions. The lowest divergence for each singular template and overall across templates is marked in bold.

Table 2.1: Mean J-S divergence by item, by singular template

	Model name	**Simple Temp Match**	**Prob. Temp Match**	**GCM, Temp-Restricted**	**Prob. GCM, Temp-Restricted**
	Fine-grained similarity	No	No	Yes	Yes
	Decision rule	Deterministic	Probabilistic	Deterministic	Probabilistic
CV Template					
CvCC		0.3846	**0.2528**	0.5450	0.2916
CvCCvC		**0.1034**	0.1937	0.1935	0.1388
CvvC		0.4295	**0.1857**	0.4050	0.1951
vCvC		0.2795	0.2730	0.2910	**0.2158**
CvCCvT		**0.0974**	0.1553	0.1783	0.1842
CvCCvvC		0.2864	**0.0557**	0.2864	0.2252
CvCvvC		0.3042	0.2685	0.3611	**0.2669**
CvCvvCvT		**0.0710**	0.1831	**0.0710**	0.1132
MEAN		0.2445	**0.1960**	0.2914	0.2038

Overall, the best-fitting model is the Probabilistic Template Match, which uses the singular template to define similarity and a probabilistic choice rule. By singular template, this

model fits best for three of the eight templates, while the deterministic Simple Template Match fits best for two templates, [CvCCvC] and [CvCCvT], and the probabilistic GCM fits best for two templates, [vCvC] and [CvCvvC]. There is also one template, [CvCvvCvT], for which the two deterministic models fit equally well[3]. The two probabilistic models show much lower mean divergences than the two deterministic models, but these sums do not entirely clarify the extent to which participants are using differing decision rules, as it appears to be the case for some singular templates but not others. The by-participant fit may clarify this better, as it will show whether this stems from a difference between participants in pluralization or from a difference that is consistent across all participants.

2.2.2.3.4 By-participant fit

Table 2.2 below shows the mean divergence between each model and the experimental data by participant. Probability distributions were calculated for each participant for each singular template, and for each model for each singular template. There were 61 participants x 8 singular templates, resulting in 488 comparisons between the experimental data and the models. The divergences for each participant were averaged for each model for each singular template (shown in rows labeled by singular template), and then averaged for each model across singular templates (shown in row labeled "mean"). The lowest divergence for each singular template and overall is marked in bold.

---

3. For the singular template [CvCvvCvT], the two deterministic models fits equally well. This is possible if they both select the same plural for all items with that singular template, as the deterministic models select the best plural rather than a probability distribution across the plurals. These two models also converged for the singular template [CvCCvvC], but not for any of the other six templates.

Table 2.2: Mean J-S divergence by participant, by singular template

	Model name	**Simple Temp Match**	**Prob. Temp Match**	**GCM, Temp-Restricted**	**Prob. GCM, Temp-Restricted**
	Fine-grained similarity	No	No	Yes	Yes
	Decision Rule	Deterministic	Probabilistic	Deterministic	Probabilistic
CV Template					
CvCC		0.4363	**0.3600**	0.7270	0.3741
CvCCvC		**0.1285**	0.2521	**0.1285**	0.1964
CvvC		0.4327	**0.2623**	0.4327	0.2654
vCvC		0.3124	0.3052	0.3124	**0.2612**
CvCCvT		**0.1077**	0.1916	**0.1077**	0.2096
CvCCvvC		0.3504	**0.1516**	0.3504	0.2951
CvCvvC		**0.3359**	0.3637	**0.3359**	0.3413
CvCvvCvT		**0.0782**	0.1970	**0.0782**	0.1238
MEAN		0.2728	0.2605	0.3091	**0.2584**

By participants overall, the best-fitting model is the *Probabilistic GCM*, although the *Probabilistic Template Match* fits nearly as well in aggregate. Of the eight templates, three singular templates fit best to the *Probabilistic Template Match*, one fits best to the *Probabilistic GCM,* [vCvC], and four fit equally well to the two deterministic models, [CvCCvC], [CvCCvT], [CvCvvC], and [CvCvvCvT]. Interestingly, the *Probabilistic Template Match* is the worst fit of all four models for [CvCCvC], [CvCvvC], and [CvCCvCvT]. The pattern of divergences by participant is similar to the pattern by item, with the exception of the [CvCvvC] template, which by item fits the *Probabilistic GCM* best, and by participant fits the deterministic *Simple Template*

Match best. In aggregate, the best model differs by item and by participant, where by item it is the *Probabilistic Template Match* and by participant it is the *Probabilistic GCM*.

It is possible that individual participants use differing strategies in plural formation, which is not determinable from the summed data above. To examine this, the best-fitting model for each participant was calculated by averaging the divergences for that participant for each model across the singular templates. Table 2.3 shows the number of participants for whom each model fits the best of the four models. Overall, roughly two-thirds of participants best fit the models using type statistics on the CV template best (*Simple Template Match* and *Probabilistic Template Match*; n=40). Interestingly, although the best-fitting models by item and by participant are the *Probabilistic Template Match* and the *Probabilistic GCM*, respectively, the single largest group is those that fit the deterministic *Simple Template Match* best (n=27). However, the number of participants fitting the two probabilistic models best is greater than the number of participants fitting the two deterministic models best (n=42 vs. n=29, respectively). This shows that some participants are choosing more probabilistically among possible plurals, while others choose more deterministically, although on average, participants select somewhat more probabilistically. Additionally, this analysis shows that some of participants do take additional fine-grained segmental features into account in determining the best analogy for a nonce form, although this is not the dominant strategy, with roughly one-third of participants best fitting one of the models using fine-grained similarity (n=21).

Table 2.3: Number of participants with best fit to each model

Decision rule	Deterministic	Probabilistic	Deterministic	Probabilistic
Fine-grained similarity	No	No	Yes	Yes
Model name	Simple Temp Match	Prob. Temp Match	GCM, Temp-Restricted	Prob. GCM, Temp-Restricted
Number of participants	**27**	13	2	19

If we examine more closely the participants who fit the deterministic *Simple Template Match* (STM) versus the participants who fit the *Probabilistic Template Match* (PTM), we find that the determinism of the first group's behavior is relative. Figure 2.5 shows the expected versus observed log probability of each plural template for each singular template, with participants divided by which of the models they fit best. Data from participants who fit the deterministic *STM* is shown on the right, and data from participants who best fit the probabilistic *PTM* is shown on the left. For *PTM*-fitting participants, r=0.71, while for *STM*-fitting participants, r=0.68. Thus, although the participants who fit the *STM* best do display somewhat more deterministic choice strategies than the participants who fit the *PTM* best, the *STM*-fitting participants nonetheless do not behave in an entirely deterministic manner, and the resulting nonce plurals still fit the expected probabilities for each plural template quite well.

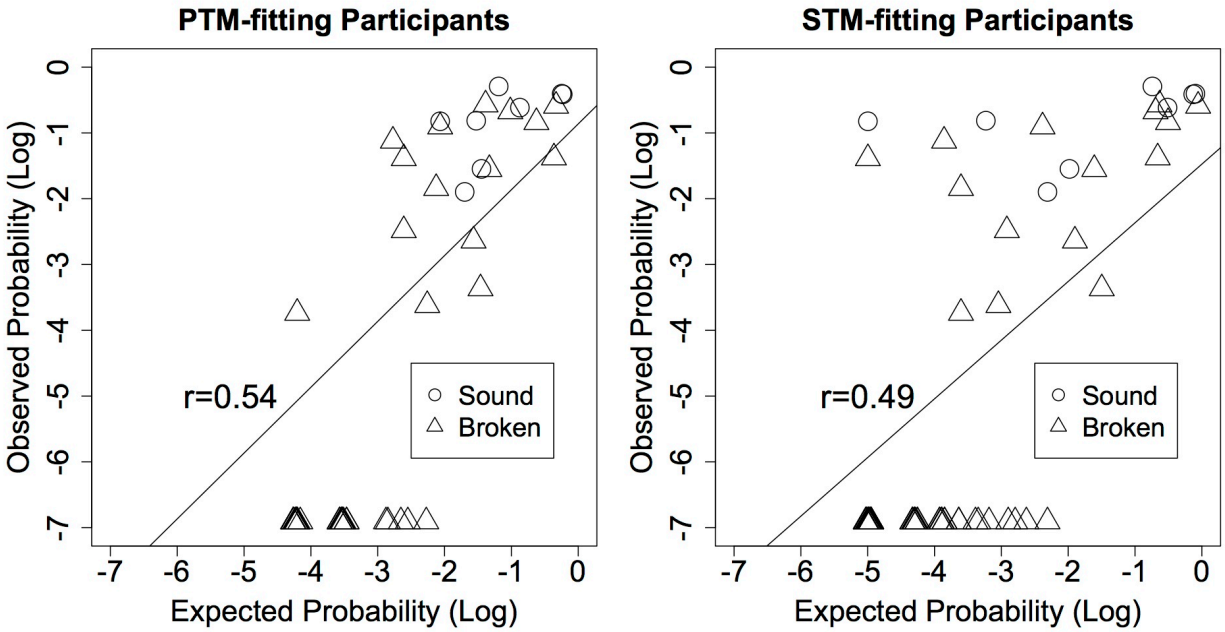

Figure 2.5: Expected vs. observed log probability of plural template by singular template, for participants fitting *Probabilistic Template Match* (left) and *Simple Template Match* (right)

The above analysis categorizes the participants by the best-fitting model, but does not examine the amount by which an individual participant fits one model better than the other. Figure 2.6 shows the divergence from the *Simple Template Match* versus the divergence from the *Probabilistic Template Match* for each participant that fit one of the two models not using segmental similarity (that is, excluding the participants who fit either the *Probabilistic GCM* or the *Template-Restricted GCM* best). If a participant is above the $x=y$ line, they fit the *STM* better, and if they are below the $x=y$ line, they fit the *PTM* better. In general, we see that participants who fit the *STM* better are closer to the $x=y$ line than participants who fit the *PTM* better, which is in concordance with the overall better fit across participants to the *PTM*. Nonetheless, the difference in divergence from the two models is quite small for many participants, and the divergences are significantly positively correlated, $r=0.59$, $p<0.05$. The difference in model fit

for a given participant is far from absolute, and indicates that individual participants generally show a tendency toward probabilistic or deterministic behavior, not that they display perfectly probabilistic or deterministic behavior.

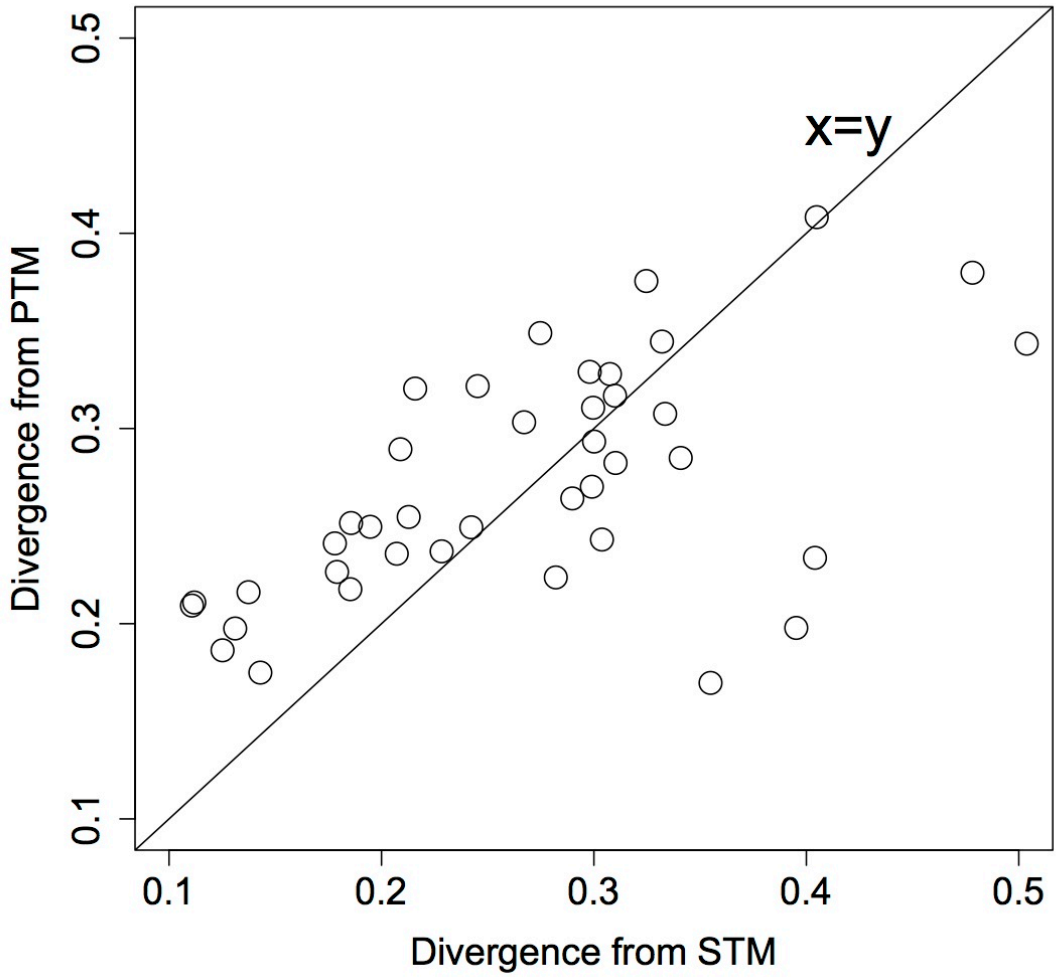

Figure 2.6: By-participant divergence from *Simple Template Match* vs. *Probabilistic Template Match*

Figure 2.7 examines the divergences in this same fashion for the two probabilistic models, the *Probabilistic GCM* (PGCM) and the *Probabilistic Template Match* (PTM), using only the participants who fit one of these two models the best. If a participant is above the $x=y$ line, they fit the *PGCM* better, and if they are below the $x=y$ line, they fit the *PTM* better. There

is a very close fit between the divergences by participant for these two models, with a significant positive correlation of $r=0.96$, $p<0.001$. This suggests that for participants who display more probabilistic behavior, the use of fine-grained segmental similarity in forming analogies creates very small differences in actual behavior in generalization.

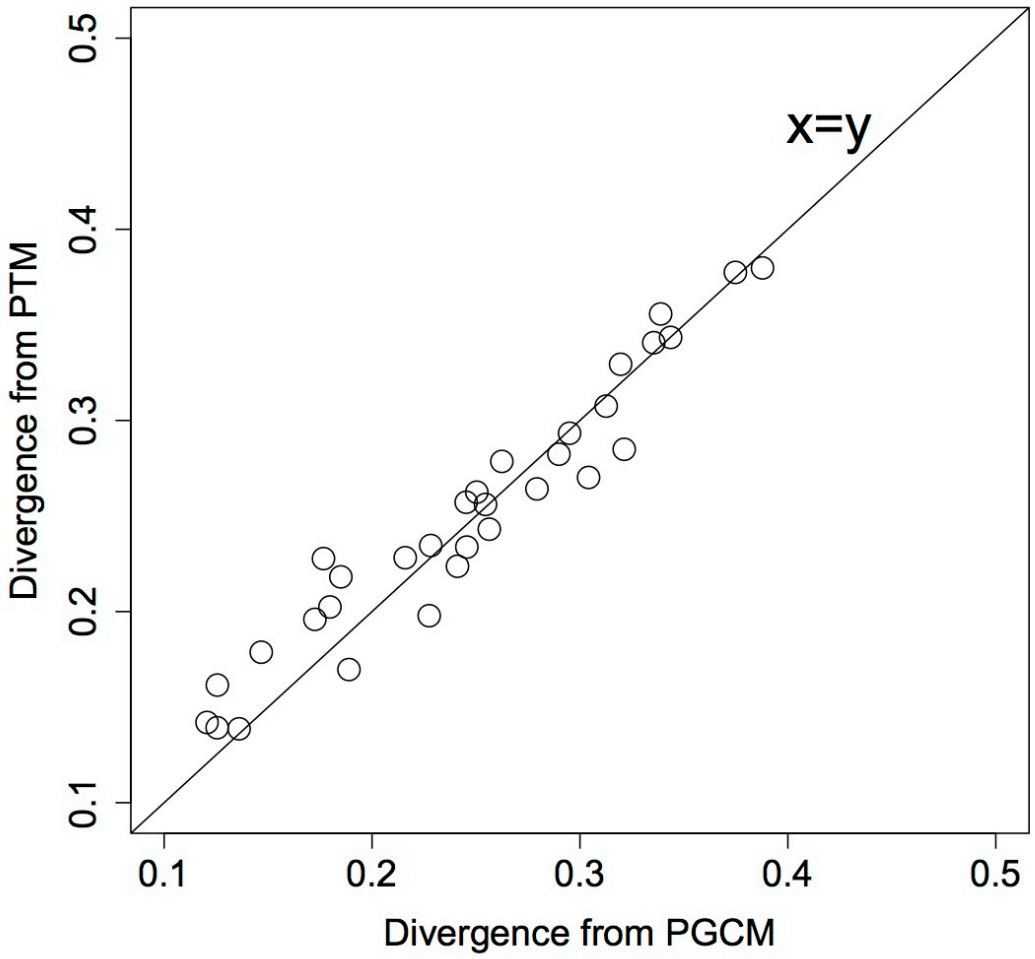

Figure 2.7: By-participant divergences from *Probabilistic GCM* vs. *Probabilistic Template Match*

One question that remains unanswered is the difference in model fit between singular templates. For four of the singular templates, the deterministic *Simple Template Match* fits participant plurals better than the *Probabilistic Template Match*, which indicates that participants

are choosing more deterministically among possible plural templates for these singular templates. In addition, for one singular template, [vCvC], the best-fitting model is the *Probabilistic GCM*, which uses fine-grained segmental similarity in additional to template structure in determining the best plural. This is the only singular template for which one of the models utilizing segmental similarity fits best[4], which suggests that there may be some difference in the informativity of segmental characteristics for items with this singular template versus items with the other singular template.

2.2.3 Discussion

Overall, this experiment demonstrates that native speakers of Arabic produce plurals for nonce singular nouns in a manner that reflect the lexical statistics of existing forms. In forming plurals for previously unseen forms, speakers rely primarily on type statistics on the CV template, which corroborates the theoretical and computational evidence that this is the primary driver of noun plural formation in Arabic (Dawdy-Hesterberg & Pierrehumbert, 2014; McCarthy & Prince, 1990a). In addition, speakers choose among the possible plurals for a given nonce singular in a probabilistic manner, producing a plural template in proportion to its frequency for that singular template in the lexicon. In a comparison to four predictive analogical models of pluralization, I showed that the experimental data best fits the models that use a probabilistic

4. Note that there are three templates for which the *Simple Template Match* and the *Temp-restricted GCM* fit best. These two models always converged on the predicted plural for all nonce items with these singular templates. Because the *Simple Template Match* does not utilize segmental similarity, this tie between the models indicates that the addition of segmental similarity is not leading to different predictions. Thus, the fact that the *Temp-restricted GCM* is tied for best fit for these templates does not necessarily indicate that participants are using fine-grained segmental similarity in deciding on the plural for nonce items with these singular templates.

choice strategy, with the *Probabilistic Template Match* fitting best by item and the *Probabilistic GCM* fitting best by participants. The model fit is extremely similar by both participant and by item, which indicates that overall participants are using similar strategies in forming plural for nonce forms.

The by-participant analysis shows interesting, and important differences in the strategies used by different participants that are not revealed by the by-item analysis. Just over half of the participants best fit a model using a probabilistic choice rule, while slightly less than half best fit a model using a deterministic choice rule. As shown in Figure 2.5, however, the amount of determinism in the participant group fitting the deterministic *Simple Template Match* is relative; the probability of a given plural in the experimental data is still strongly correlated with the expected probability based on lexical statistics. This suggests that all participants are nonetheless sensitive to type statistics, although some participants tend toward more, though not entirely, deterministic strategies in selecting among possible plural templates, a pattern which is reiterated in Figure 2.6. This finding mirrors some results in artificial language learning, where individual differences in the tendency to probability-match versus over- or under-regularize in generalization arise when there is high variability in the input (e.g., Hudson Kam & Newport, 2005; Schumacher et al., 2014). The split between probability-matching and deterministic behavior is also displayed in the model fit by singular templates. For three of the singular templates, the best-fitting model by item and by participant is the *Probabilistic Template Match*, while the deterministic version of this model, the *Simple Template Match,* fits three templates best by item and four templates best by participant.

Interestingly, within the probabilistic models, there is also a split in participants in terms of use of fine-grained segmental similarity. While this split in model fit also appears in the

aggregate fit by item and by participant, the use of additional fine-grained segmental similarity in analogy formation makes only a small contribution to the plurals given by participants. Participants who best fit the *Probabilistic GCM* and the *Probabilistic Template Match* show extremely similar distributions of plurals, as shown by the strong correlation between divergences for these two models in Figure 2.7. This mirrors the modeling results in Dawdy-Hesterberg & Pierrehumbert (2014), where the model using fine-grained segmental similarity in addition to type statistics on the CV template (Template-restricted GCM) performed significantly, but only slightly, better in predicting plurals for unseen forms that the model using only type statistics on the CV template (Simple Template Match), with an increase in accuracy of about 2%. As to why some participants seem to utilize fine-grained segmental similarity, while others don't, this is an open question. The participants fitting these two models saw roughly equal numbers of each set of nonce items, so this difference could not have stemmed from the specific items encountered by each group. Nonetheless, given the current data, we can conclude that the CV template of the singular has a demonstrable effect on speaker behavior in generalizing plurals to new forms, while fine-grained segmental similarity has only a weak (if any) effect on the behavior of some participants.

One unanswered question is whether speakers have strong preferences for the plural for an unseen form. Even if speakers are uncertain about the outcome in an open-response paradigm, as evidenced by the variability in responses observed in experiment 1A, they may nonetheless show preferences for particular plurals when the plurals are given to participants. Experiment 1B will examine this using the same nonce forms from experiment 1A in a forced-choice paradigm. By presenting two possible plurals for an unseen form, we can examine speaker preference for the possible plurals. This paradigm will answer whether the uncertainty about the outcome is

partially a factor of the open-response paradigm, or if it is a result of the uncertainty in the system as a whole.

In addition, as noted, there is some variability in the number and type of plurals given by participants for nonce items with the same singular template. Because of the limitations of the models, as well as the relatively weak effects of segmental similarity in Dawdy-Hesterberg & Pierrehumbert and the current experiment, we cannot say definitively that the participants who best fit models that do not use fine-grained segmental similarity are not using the specific segmental characteristics of each nonce item at all in experiment 1A. If participants in general are using segmental characteristics in analogy formation in experiment 1A, albeit to varying degrees, then speaker responses overall should be better predicted by the probabilities of the responses in experiment 1A for that specific nonce item than the probabilities of those plurals in the corpus dataset, which is calculated only by the singular CV template. Thus, the follow-up experiment will give more insight into the extent to which plural preferences differ across nonce items, and the relative strength of specific segmental characteristics in analogy formation for plurals.

2.3 Experiment 1B

2.3.1 Introduction

This experiment uses the stimuli and responses from experiment 1A in a forced-choice task to examine if and how participants display preferences for particular plurals for the nonce singulars from experiment 1A. As demonstrated above, in an open response task, participants in aggregate do not choose deterministically, but rather produce plurals in manner that reflects their

statistical distribution in the lexicon. This may be due to high uncertainty about what the outcome should be, or to the higher cognitive load necessary to generate a plural when the options are unconstrained. In a forced-choice paradigm, however, participants are given a finite set of options, so the search space is constrained, and speakers may select among possible choices differently than in the unconstrained open-response paradigm.

Moreover, we have item-specific probabilities for the plural templates from the responses in experiment 1A. Although the modeling work in Dawdy-Hesterberg & Pierrehumbert (2014) demonstrated that the specific segmental features of the singular added only a little predictive power in an analogical model, and the model fit from the previous section showed that, by item, a probabilistic model that did not use segmental features in forming analogies was a better fit to the experimental data than a similar model using the segmental features, nonce items with the same singular template in some cases showed very different distributions of plurals in experiment 1A. Thus, it is possible that the number of items and/or subjects in experiment 1A was simply too small to detect the influence of segmental features on plural formation when calculated by item.

In this experiment, participants were given a forced choice between two plurals given by participants in experiment 1A. A total of three possible plural choices were examined, with each participant seeing one of the three possible pairs. By using the specific plurals for each nonce item produced in experiment 1A, we can examine the extent to which the probability of a plural for a singular template, and the probability of a plural for individual nonce items in experiment 1A, predict which plural choice speakers prefer. This will allow us to examine in more detail the extent to which there are item-specific differences that stem from differences in segmental features beyond those defined by the CV template.

2.3.2 Methodology

2.3.2.1 Participants

Participants were recruited via Amazon's Mechanical Turk. The heading for the experiment was "Answer a survey about Arabic words" (in Arabic). Participants received $4 upon completion of the experiment. Participants who took part in experiment 1A were blocked from participating in experiment 1B. In total, 381 participants accepted the task on Mechanical Turk, of which only 151 completed any experimental items. 135 participants completed the entire experiment. 68 of those participants were excluded from analysis for the following reasons: database error resulting in demographic information not being recorded (n=9), having completed experiment 1A previously (n=5), not being a native speaker of Arabic (n=32) or achieving less than 80% accuracy on filler items (n=22). In total, 67 participants completed the experiment and met all qualifications for inclusion in analyses.

Of the 67 participants whose data was analyzed, 43 participants were male and 18 were female. Gender was not recorded for 6 participants due to a database error; all other demographic information was recorded for those participants so the participants were included in analyses**.** All analyzed participants were self-reported native speakers of Arabic. Mean proficiency in MSA was 8.68 on a scale of 1-10 (S.D.=1.59). Mean frequency of use of MSA was 6.61 on a scale of 1-10 (S.D.=2.51), with 1 being "rarely use MSA" and 10 being "use MSA frequently." For level of education, 7 participants reported having less than a college education, 47 participants an undergraduate education, 8 participants a master's degree, and 5 participants a doctorate. 62 participants reported also speaking English, with a mean proficiency of 8.27 (S.D.=1.87) on a scale of 1-10. All but 2 participants reported speaking a second language (including English), 25 reported speaking a third, and 6 reported speaking a fourth.

Information on primary spoken dialect was also elicited. Dialects were classified by major regional dialect. 16 participants reported speaking Egyptian as their primary dialect. 18 participants reported speaking a Levantine dialect (includes Jordanian, Syrian, Palestinian & Lebanese). 14 participants reported speaking a North African dialect (includes Moroccan, Sudanese, & Tunisian). 8 participants reported speaking a Peninsular dialect (includes Kuwaiti, Saudi, Emirati, & Hijazi). 5 participants reported speaking a Mesopotamian dialect (all Iraqi), and 6 participants did not specify a dialect.

2.3.2.2 Experimental materials

2.3.2.2.1 Stimulus design

This experiment examines the same eight singular templates as experiment 1A. The same 48 filler items were used for experiment 1B, which were all moderate-to-high frequency existing singular nouns that take the dominant plural template for that singular template. The nonce items were three of the five sets of 48 nonce items from experiment 1A, for a total of 148 nonce items.[5] Each participant saw one of three sets, with the sets counterbalanced across participants. For each item, the participant saw two possible plurals and selected the plural they preferred. For the filler items, all participants saw the real plural and a distractor. For the nonce items, each participant saw one of three pairs of nonce plurals, which were selected from the responses to experiment 1A. For each nonce item, the three most-frequent response patterns were selected, such that the following criteria were met:

5. Only three of the five sets from experiment 1A were examined in this task, as the two-way forced-choice paradigm using three possible plurals required three times the number of participants. Thus, to limit the overall number of participants, only a subset of the nonce items from experiment 1A were examined.

1. For each nonce item, no selected plurals could share the same CV template. That is, if [CiCaaC] and [CuCuuC] were both among the three most frequent plural patterns, only the more frequent of these in experiment 1A was selected.
2. Any responses taking the sound plural were counted as the same response template, regardless of variation in short vowel insertion (e.g., CaCC⇒CaCaC+aat) and CaCC⇒CaCC+aat were considered the same response template)[6]. The actual stimulus presented to participants was the variant that was more frequently given in experiment 1A.
3. If two patterns had the same frequency for a given nonce item, the one which was more frequent across all nonce items with the same singular template was selected.
4. If there were fewer than three patterns with distinct CV templates among the responses, the most-frequent pattern for that singular template that was not given for that nonce item were used to fill the remaining pattern slot(s).

This heuristic preserved the variation seen across nonce items, as individual items with the same CV template did not necessarily show the same ranking of plurals in experiment 1A. In addition, this heuristic ensured that participants selected between different plural templates, as that is the primary dimension under investigation, while ensuring that the vowel patterns in the presented forms were not anomalous. Future experiments may wish to examine the vowel quality of plurals in more detail; however, previous work has shown it to be a secondary factor in

6. This method was employed because the main question under examination was the selection among CV templates. The previous modeling work considered any [-aat] suffixation to be one template, so this heuristic was preserved in this experiment as well. Although the question of short vowel insertion in [-aat] suffixation is interesting, and has been examined to some extent (e.g., Ratcliffe, 1998), the current experiment is neither focused on this question nor able to evaluate it adequately.

pluralization (Dawdy-Hesterberg & Pierrehumbert, 2014; Ratcliffe, 1998), and thus these experiments focused on the CV template.

Each participant saw two of the three plurals for a given item (A vs. B, B vs. C, or A vs. C). This was counterbalanced across participants such that an equal number of participants saw each possible combination. Responses were coded according to the number of times a participant in experiment 1A responded with the selected CV template, such that the plural with the most-frequent CV template for that nonce item was A, the second-most frequent was B, and the third-most frequent was C.

For the filler items, all participants saw the same pair: the attested plural, and a plural constructed with the next most-frequent plural pattern for that singular CV template from experiment 1A. For example, for the filler item [maktab], participants chose between [makaatib] (the correct plural) and [maktabaat] (a distractor using the most common plural pattern after [CaCaaCiC] for that singular template). Thus, the filler distractors all had CV template structure that was attested for that singular CV template in experiment 1A.

2.3.2.2.2 Procedure

The procedure was identical to experiment 1A, except that participants were given two choices for the plural and asked to select which form they preferred. Figure 2.8 shows an example page. The experimental materials from experiment 1A were used, with slightly modified instructions and example screens to accommodate the forced-choice paradigm rather than the open response paradigm. The same filler items were used, and were also as qualifying

questions with the same threshold of 80% accuracy. Analyzed participants had a mean accuracy of 94.1% (S.D.=0.46).

In addition to sentence frame counterbalancing, and randomization of order of presentation, the two plural options were counterbalanced for order.

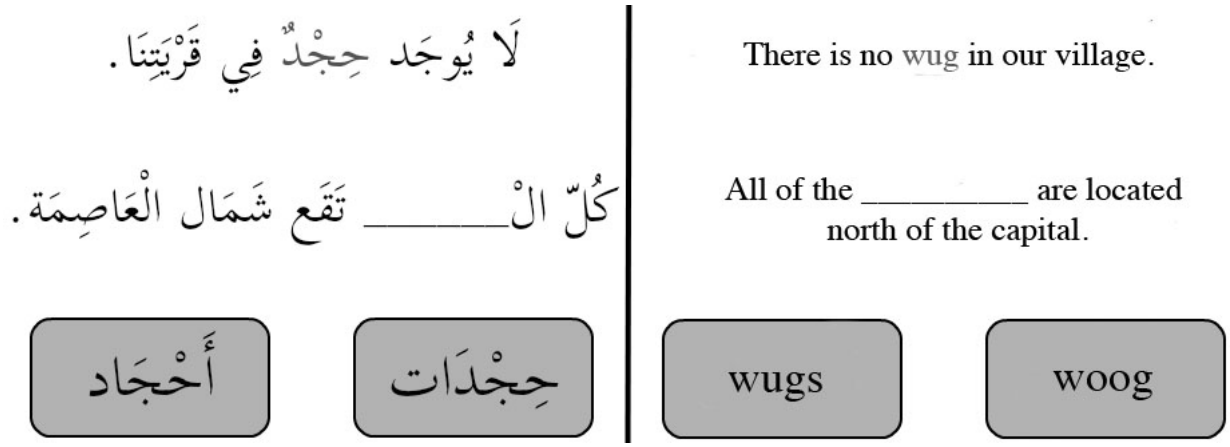

Figure 2.8: Example nonce item from experiment 1B (left) and English gloss (right)

2.3.3 Results

2.3.3.1 Overall results

Overall, we find that participants do prefer the higher-ranked plural. Figure 2.9 shows the proportion of responses for each plural option, where "A" is the most-frequent plural for that item from experiment 1A, "B" is the second-most-frequent, and "C" is the third-most frequent (following the stimulus design procedure outlined in the methodology). In aggregate, participants show a preference for plural A over plural B, selecting A 56.8% of the time. Participants show a slightly stronger preference for plural A over plural C, selecting A 70.0% of the time. Finally, participants show a slight preference for plural B over plural C, selecting B 51.2% of the time.

As expected from the results of experiment 1A, participants do not select deterministically, but the results below show that there is a general preference for the more-frequent plural.

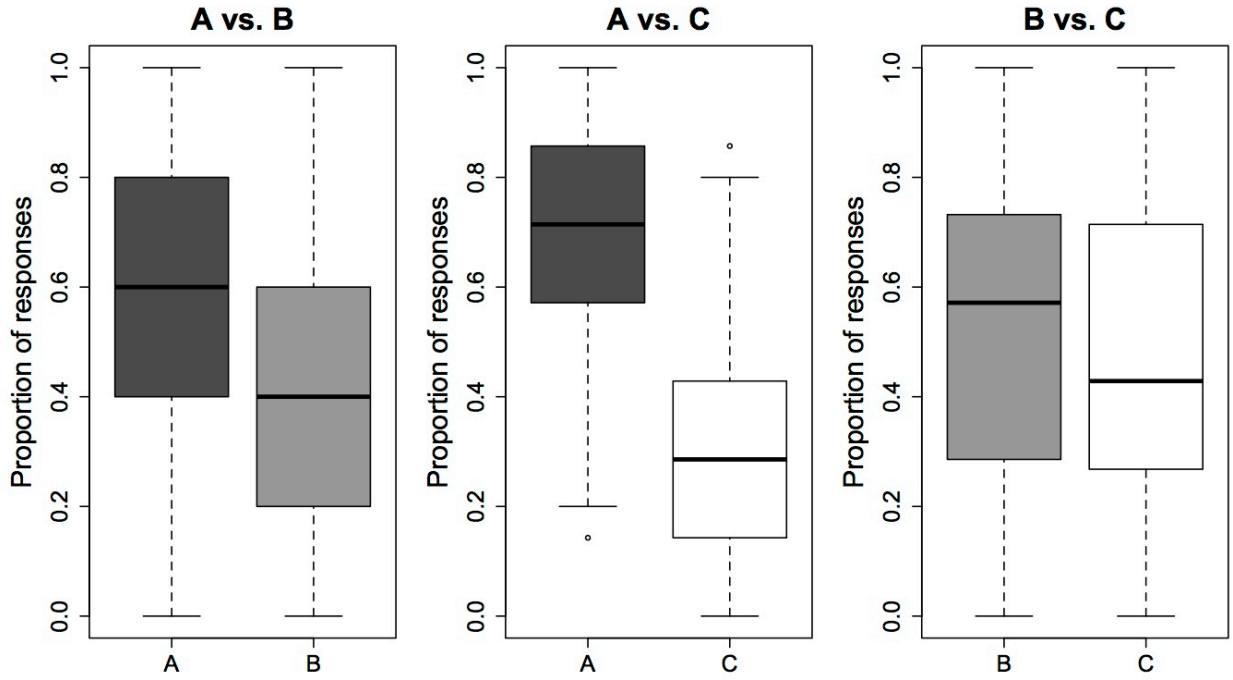

Figure 2.9: Proportion of responses for plural templates by ranking

However, this pattern varies quite a bit across individual items, and across singular templates. For some items, 100% of participants selected A over B, while for others, there is a 50-50 split. Likewise, for singular templates, there is quite a bit of variability. Figure 2.10 shows the proportion of responses by singular template, calculated by item. For some singular templates, the preference for A over B is nearly absolute, for instance the [CvCCvC] items, while for others such as [CvCvvC] items, there actually appears to be a preference for B over A. There are a number of reasons why this might be the case. First, the number of participants in experiment 1A was relatively small, with an average of 12 participants giving a judgment on any one item. Thus, the probability estimates for the output plurals are relatively unstable, and could

certainly have been bolstered by additional participants. Although the experimental probabilities were strongly correlated with corpus probabilities, there is certainly more variation in the experimental probabilities due to the number of participants. Second, the type frequencies of particular plurals, and the difference between these frequencies, varies quite a bit across gangs. However, the individual plural templates, and the ranking of the plural templates, varied quite a bit across items. Thus, the aggregate results may not tell the whole story about where the variability in preference rankings stems from.

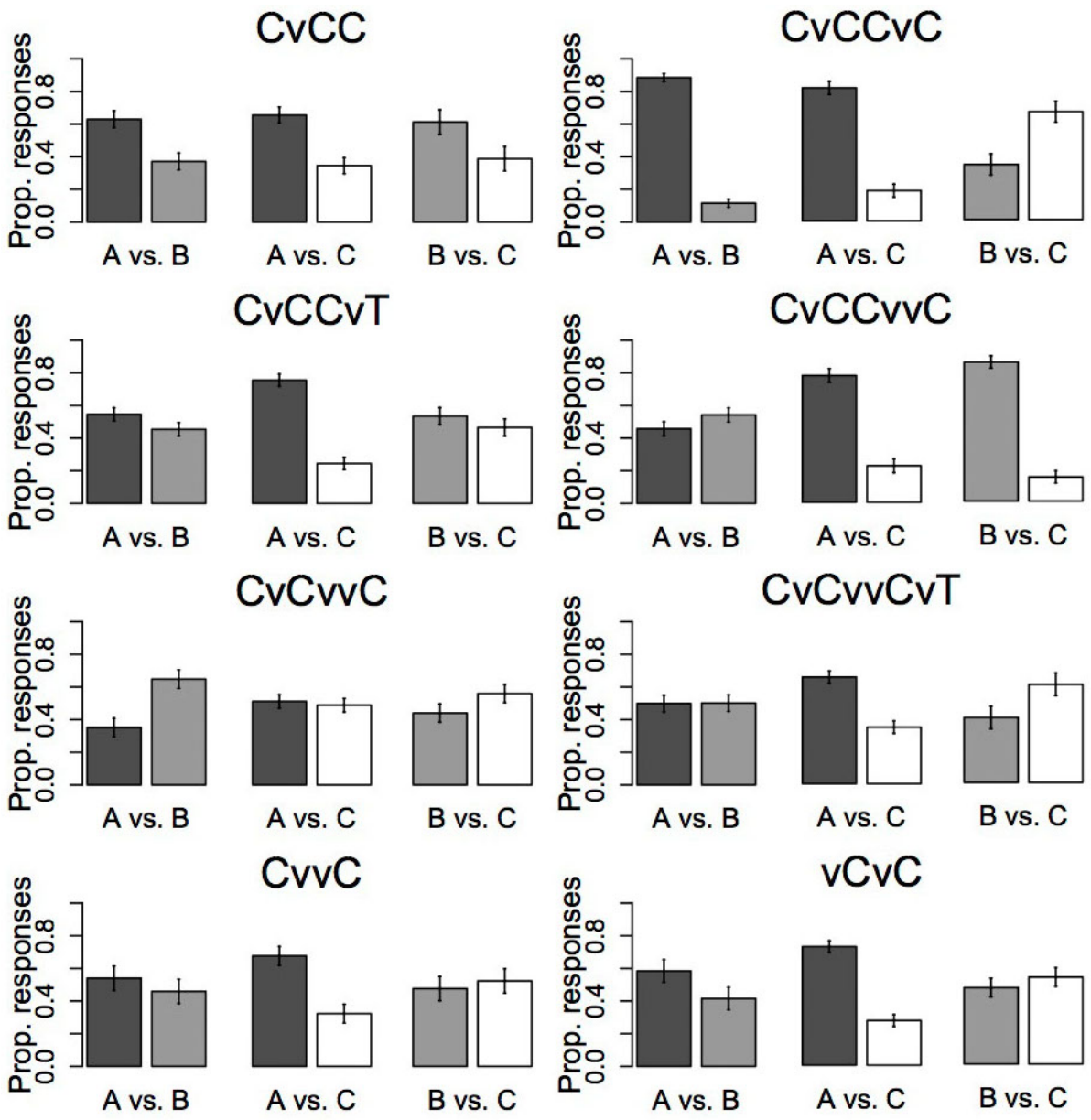

Figure 2.10: Proportion of responses by item for plural templates by ranking, by singular CV template (error bars show S.E.)

As noted, there are two different sets of probabilities for each plural template for each item. First, there are the probabilities of each plural for each singular template taken from Dawdy-Hesterberg & Pierrehumbert (2014). If you recall, the model using these probabilities and a probabilistic choice rule was the best fit to the experimental data by item in experiment 1A. Second, there are the probabilities for each plural template for each item taken from the

experiment 1A results. Thus, we now wish to assess whether participants in the forced-choice experiment are also basing their preferences on the overall probabilities, as the majority of participants seemed to be doing in experiment 1A, or whether the item-specific probabilities from the results of experiment 1A are better estimates of participant preferences.

In order to assess the effect of the corpus probabilities of the various plural templates versus the effects of the probabilities of the plural templates from the responses of experiment 1A, a linear mixed-effects regression was used. As noted, the corpus probabilities used were those from Dawdy-Hesterberg & Pierrehumbert, and the experiment probabilities were drawn from the results of experiment 1A. A model was constructed which tried to predict whether a participant would select the higher-ranked plural for a particular item using the following fixed effects: corpus probability of the higher-ranked plural, corpus probability of the lower-ranked plural, probability of the higher-ranked plural in experiment 1A, probability of the lower-ranked plural in experiment 1A, and which version of the experiment the participant saw (where version 1 is plural A vs. plural B, version 2 is plural A vs. plural C, and version 3 is plural B vs. plural C). In addition, the model included a random intercept by participant. No interactions between factors were included. Significance for each factor was determined used nested model comparison (Barr, Levy, Scheepers, & Tily, 2013).

First, we find that the experiment 1A probabilities of the higher- and lower-ranked plurals are both significant predictors of whether a participant will select the higher-ranked plural. The probability of the higher-ranked plural is a significant positive predictor, β=1.24, S.E.=0.24, $\chi^2(1)$=26.32, p<0.001. The probability of the lower-ranked plural is a significant negative predictor, β=-2.02, S.E.=0.45, $\chi^2(1)$=19.66, p<0.001. However, neither the corpus probability of the higher- or the lower-ranked plural is a significant predictor of whether the participant will

select the higher-ranked plural, β=-0.23, S.E.=0.23, $\chi^2(1)$=-0.99, p=0.32 and β=0.28, S.E.=0.23, $\chi^2(1)$=1.48, p=0.22, respectively. Finally, version is not a significant predictor of whether the participant will select the higher-ranked plural, β=0.03, S.E.=0.09, $\chi^2(1)$=0.11, p=0.74.

2.3.4 Discussion

This experiment demonstrates that participant preference for the plurals of nonce singular nouns is generally in line with the responses from the open-response experiment 1A. That is, participants show a general preference for the most-frequent response (plural A) over the second most frequent response (plural B) and the third most frequent response (plural C). Likewise, participants show a slight preference for plural B over plural C. The preferences overall are not categorical. This not-deterministic behavior is expected given the results in experiment 1A and similar literature on nonce-form generalization in morphological systems with high uncertainty. Overall, though, participants show a general preference for the more-frequent plural out of a given part of plural responses, and the strength of the preference is related to the overall ranking of the plural. Participants show the strongest preference when given the choice between plural A and plural C, while participant preferences are weaker for plural A versus plural B, and weakest for plural B versus plural C.

When we examine the preferences by singular template, we see some interesting differences across the templates. For all 8 singular templates, plural A is preferred over plural C, although the strength of the preference varies quite a bit across the templates. For instance, for the template [CvCCvC], the preference is over 80% for plural A, while for the template [CvvC],

the preference for plural A is just over 50%. Interestingly, there are also some gangs for which the preference for A over B or B over C is reversed. Participants show a slight preference for plural B over plural A for the templates [CvCCvvC] and [CvCvvC]. In addition, participants show a preference for plural C over plural B for five templates, although this preference is very slight for three of these templates. A major question is why there are such differences across the singular templates. One explanation is that the number of licit existing plural templates varies across singular templates. For example, for the singular template [CvCCvvC], there are only two extant plural templates: [CvCCvvC + aat] and the broken plural [CvCvvCvvC]. In experiment 1A, participants entered some plural templates that never occur in the lexicon for that singular template, although they are existing plural templates for other singular templates. For nonce items with this singular template, in fact, the third-ranked option for every single item was a plural template that does not occur in the lexicon with that singular template. Thus, it is not surprising that participants show a strong dispreference for plural C for items with this template. In contrast, for the singular [CvCvvC], there are eight existing plural templates in the dataset from Dawdy-Hesterberg and Pierrehumbert. The specific ranking of plurals varies quite a bit across nonce items, which may partially explain the very different preference pattern for this singular template. The overall differences across templates mirrors differences observed in experiment 1A, which are at least partially explained by the fact that different singular templates have varying numbers and probabilities of plural templates in the lexicon.

The mixed-effect model directly compared the effect of the probability of a plural template for that singular template with the probability of a plural template for the specific nonce item in experiment 1A. The overall results are somewhat surprising, as the model found that only the probabilities of the plural template for that item in experiment 1A were statistical predictors

of participant preference, and that the probabilities of the plural template by singular template in the lexicon were not statistically significant predictors of participant preference. In experiment 1A, the corpus probabilities of a plural template by singular were strongly correlated with the response probabilities. In addition, the model using these corpus probabilities and a probabilistic choice rule was the best-fitting model of the four when examined by item. Thus, it is surprising that these probabilities have no significant effect at all in this experiment when other factors are controlled for. As previously noted, there was no overlap in participants between the two experiments, so this is not a possible explanation for this result.

There are a few possible explanations for this finding. First, it is possible that the segmental characteristics of the nonce singular do have an effect for the majority of participants on pluralization, but that this effect was small enough that it was only to detectable for a small proportion of participants in experiment 1A. There are a few supporting pieces of evidence for this theory. In the model comparison in experiment 1A, the difference in fit between the two probabilistic models, where one model used fine-grained segmental similarity in assessing analogy, and the other used only type frequency on the singular template in assessing analogy, was quite small, which indicates that the model using segmental similarity in assessing similarity fit nearly as well as the one which used only the CV template in assessing similarity. This theory is also supported by the fact that the model using segmental similarity was the best fitting of the models when calculated by participant. In addition, there were many differences in the rankings of the plurals for individual items within some gangs, which could be attributed to the segmental characteristics of the particular nonce item. Given a larger number of participants and/or items, it is possible that the effects of segmental similarity that were observed in the modeling work in

Dawdy-Hesterberg and Pierrehumbert (2014) could also be observed experimentally for a larger number of participants in an open-response paradigm like experiment 1A.

The second possibility is that the corpus estimates collected in Dawdy-Hesterberg & Pierrehumbert were somewhat unstable. The cross-validation protocol used in the paper required 4 items with the same singular template taking the same plural template, and thus a number of lower-frequency plurals were not included in the dataset. This led to a relatively large number of plural responses in experiment 1A for which there was no corpus probability estimate. Moreover, the corpus from which the dataset was collected was relatively small at roughly 850k words, so a larger corpus may have uncovered more of the lower-frequency plurals. Unfortunately, there are few large corpora available for Arabic, and none to my knowledge that contain multiple genres of text, so compromises in the type of corpus used must be made. In any case, we must always treat frequency data as unstable estimates, not as strict indicators of likelihood.

2.4 General Discussion

In sum, these two experiments demonstrate that native speakers of Arabic track statistics in the lexicon for noun plurals, and reproduce these statistics in generalization to unseen forms. Experiment 1A demonstrates that in an open-response paradigm, native speakers of Arabic generalize existing plurals to nonce forms in a manner that reflects the statistics of the existing plurals for each singular CV template. In addition, experiment 1A demonstrates that speakers in aggregate use a probability-matching strategy in deciding amongst possible plurals, even when there are as many as eight possible plural templates. Further, this experiment demonstrates that the primary representational level on which lexical statistics are tracked is the CV template.

Experiment 1B examines speaker preference for particular plurals from the responses given by participants in experiment 1A. This experiment finds that overall, there is gradient preference for the three most frequently input forms for a nonce singular noun, where speakers generally prefer the most-frequent plural from experiment 1A over the next two most frequent choices, and also generally prefer the second most frequent plural over the third most frequent plural. These preferences, like the responses in experiment 1A, are far from categorical, indicating that participants generally use a probability-matching strategy in both open-response and forced-choice tasks when faced with uncertainty about the optimal choice for the plural of a previously unseen item.

The statistical analysis of the results of experiment 1B compares the statistical predictiveness of the lexical probabilities of the plural template for a given singular template and the probabilities of the plural templates for the specific nonce item in experiment 1A, and finds that only the latter is a statistically significant predictor of which plural a participant will prefer. This seems contradictory to the results of experiment 1A, where lexical probabilities were strongly correlated with observed probabilities for plurals. However, as noted in the previous discussion, this discrepancy suggests that speakers may rely more heavily on fine-grained segmental similarity than either the results of experiment 1A or the modeling work in Dawdy-Hesterberg & Pierrehumbert would indicate. Critically, participants in experiment 1B saw fully diacritized forms, and thus had access to all vowel qualities for short vowels, whereas the models in experiment 1A used undiacritized forms (for methodological reasons explained in Dawdy-Hesterberg & Pierrehumbert). Thus, the probabilities on which participants are drawing in deciding amongst two possible plurals may not be lexical probabilities calculated strictly on the CV template, but rather probabilities that incorporate both the probabilities of the plural template

given the singular CV template as well as the fine-grained similarity to existing forms. This more closely mirrors the best-performing model from Dawdy-Hesterberg and Pierrehumbert, which incorporated both type statistics on the CV template and fine-grained segmental similarity to existing forms in determining the plural for an unseen word. If the model comparison in experiment 1A used fully-diacritized forms, then we would expect to see a larger effect of fine-grained segmental similarity on nonce-form generalization. In addition, given the small size of the corpus from which the probabilities used by the models were drawn, the results of experiment 1B suggest that speakers have stronger and more detailed statistical knowledge of existing forms than a relatively small corpus can capture, as a native adult speaker has certainly been exposed to more than 850k word tokens in their lifetime.

Overall, the results of these experiments corroborate the theoretical and computational evidence that the primary driver of noun plural formation in Arabic is the CV template of the singular (Dawdy-Hesterberg & Pierrehumbert, 2014; McCarthy, 1981). Importantly, the primary determinant of the plural template for an unseen singular noun is type statistics on existing forms on the level of the CV template, which is a coarse-grained generalization. The results of experiment 1B also suggest that fine-grained segmental similarity to existing forms may play a greater role in noun pluralization than estimated in Dawdy-Hesterberg & Pierrehumbert and in experiment 1A. Overall, this indicates an interesting contrast in the use of both coarse- and fine-grained similarity in forming analogies to existing words. In the noun plural system, the coarse-grained similarity is defined by the shared CV template of the singular, while the fine-grained similarity is defined by any additional shared segmental features beyond those indicated by the CV template. This system provides a contrast to many non-concatenative morphological

systems, where shared segmental similarity is the primary driving force in analogy formation (e.g., Alegre & Gordon, 1999; Ernestus & Baayen, 2003).

In addition, these results show that speakers of Arabic in aggregate use a probabilistic decision rule in deciding among possible plurals for an unseen form. The tendency toward probability-matching in systems with high uncertainty has been previously demonstrated in the morphological and morphosyntactic literature. However, there are three aspects of this work that differ from much of the previous literature on probability-matching. First, the task is a natural-language one in which native speakers track lexical frequencies in their own L1, whereas a great deal of the literature on probability-matching has used artificial language paradigms (e.g., Culbertson & Smolensky, 2012; Hudson Kam & Newport, 2005; Schumacher et al., 2014) or second-language learners (Walter, 2011).

Second, in most of the previous work demonstrating probability-matching, the possible outputs are binary (e.g., Ernestus & Baayen, 2003; Hayes et al., 2009; Hudson Kam & Newport, 2005; Schumacher et al., 2014), whereas in this work, there are a large number of possible outputs. Even if speakers in Arabic are restricting their possible plural choices to those that occur with the same singular template, there are as many as eight plural templates for some singular templates. This provides critical evidence that speakers are able to track a large number of possible outcomes and, moreover, do use the majority of them in generalizing to unseen forms. Similar work on probability-matching in uncertain systems suggests that the lower-frequency cases should drop out (Culbertson & Smolensky, 2012; Culbertson, Smolensky, & Legendre, 2012; Hudson Kam & Newport, 2009), but these results show that speakers are quite willing to use both high- and low-frequency patterns. Walter (2011) did previously demonstrate that Arabic learners matched lexical probabilities for six classes of plurals in forming plurals for existing

nouns, but critically, the probabilities were calculated across the entire lexicon, not on the singular CV template, which brings me to the final point of novelty in the current study.

Finally, the representational level on which speakers track lexical probabilities differs in this work than in previous work. There has been demonstrated probability-matching for a number of single-feature alternations, including phoneme alternations in Dutch (Ernestus & Baayen, 2003), vowel harmony alternations in Turkish (Hayes et al., 2009), and article alternations in an artificial language paradigm (Hudson Kam & Newport, 2005). In this experiment, participants track lexical probabilities on a coarse-grained phonological representation, the CV template, which cannot be defined by a single feature or alternation, but rather by the skeletal structure of the entire word, and use this in conjunction with fine-grained segmental similarity to determine the best possible plurals for an unseen singular. This indicates that the representational level of the CV template is active in morphological processing. There is evidence for the psychological reality of the CV template from psycholinguistic experiments in Arabic, where words are primed by forms sharing the CV template even when no other features overlap (Boudelaa & Marslen-Wilson, 2004). The current experiment corroborates this, and demonstrates that speakers also maintain knowledge of lexical probabilities on this representational level.

In addition, we find individual differences among speakers in the tendency to use a probability-matching decision rule versus a deterministic decision rule. The speakers that use a probabilistic strategy do not differ significantly from those that use a more deterministic strategy in any obvious way, as there seems to be no consistent effect of dialect on pluralization, nor is there a statistical difference between the two groups in their accuracy on filler items. As noted, there is some work showing similar individual differences in artificial language learning when

speakers are presented with high variability in the input (Hudson Kam & Newport, 2005; Schumacher et al., 2014), where some speakers tend toward probability-matching and others tend toward over- or under-regularization, but the possible source of these differences has not been established. This is certainly a relevant and important area for future study in language learning.

In sum, these studies exemplify the tendency toward coarse-grained generalization when such statistical generalizations are necessary to capture the morphological system. Importantly, the ability of a speaker to generalize across word types on a coarse-grained level, such as the CV template, does not rule out the ability to also use fine-grained, word-specific segmental information in determining the best analogy for an unseen form. These experiments corroborate the computational evidence from Dawdy-Hesterberg & Pierrehumbert that it is this intersection of coarse-grained abstraction across existing words and fine-grained similarity to existing words that drives generalization for the Arabic noun plural. In addition, these studies demonstrate that speakers can, and do, exhibit probability-matching behavior in generalization in a system when this probability-matching relies on tracking lexical statistics on this coarse-grained generalization and concurrently calculating similarity to existing words on a fine-grained segmental level.

Chapter 3 : Statistical regularities in the Arabic masdar system

3.1 Introduction

The Arabic verbal noun (henceforth masdar) system for form I (underived) verbs is relatively understudied, yet offers important insight into general principles of morphophonological pattern-learning in Arabic. First, there are a large number of patterns available in the system, with as many as 44 patterns cited in classic grammars (Wright, 1988). Second, the range of potential cues to masdar pattern that have been noted in the literature is large, including phonological features, transitivity, and semantics of the verb (Ryding, 2006; Wright, 1988). Third, there has been little if any study of the statistical utility of these cues in predicting masdar form, as the traditional grammars merely point to features as potentially relevant, but provide no larger examination of the system as a whole.

The masdar in Arabic is similar to the gerund in English, where the masdar is the nominal form denoting the action of the verb. In general, the masdar indicates an untensed state of the verb without reference to a subject or object, as in "I like running," but it can also indicate a single instance of performing the verb as in "His acting in the play last night was emotive" (e.g., Grenat, 1996; Wright, 1988). The potential predictive cues to the masdar form of form I verbs span a wide range of linguistic levels. The traditional analyses point to phonological, syntactic and semantic cues as relevant to masdar formation. The wide range of factors pointed to in the literature is intriguing, as the combination of these disparate types of factors presents an issue for learnability. Further, to my knowledge, the only existing analyses of the masdars of form I verbs use traditional methods of analysis, and do not examine the statistical utility of these cues in predicting masdar form. This chapter uses a large set of verb-masdar pairs from a dictionary in conjunction with text corpora to examine what, if any, statistical regularities exist in the system.

Specifically, this chapter focuses on the utility of phonological, syntactic, and semantic properties of the verb in predicting masdar form. The extent to which the factors identified by these analyses are used by native speakers in learning verb-masdar correspondences and in generalizing to new forms will be examined in the next chapter.

In this and the following chapter, I focus on a subset of Arabic verbs, the so-called form I verbs. Form I is the basic, underived form of the verb, from which the other classes in the verbal paradigm are derived. There are ten commonly used verb forms (also sometimes referred to as "measures"). The masdar of the derived verbs (forms II+) are highly regular, with form II and III verbs taking two main patterns for each form, and forms IV+ each taking one pattern with rare exception (e.g., Grenat, 1996; Wright, 1988). Table 3.1 shows the ten verb forms and their masdars, each shown with the verbal root [f ʕ l].

Table 3.1: Verb form I-X patterns and masdars

Verb form	Verb pattern	Masdar
I	faʕala/faʕila/faʕula	Various
II	faʕʕala	tafʕiil, tafʕilaT
III	faaʕala	mufaaʕalaT, fiʕaal
IV	ʔafʕala	ʔifʕaal
V	tafaʕʕala	tafaʕʕul
VI	tafaaʕala	mutafaaʕalaT
VII	ʔinfaʕala	ʔinfiʕaal
VIII	ʔiftaʕala	ʔiftiʕaal
IX	ʔifʕalla	ʔifʕalaal
X	ʔistafʕala	ʔistifʕaal

Form I verbs show great variety in masdar form, and in fact have been cited in the literature as being highly or entirely unpredictable (Grenat, 1996; Holes, 2004; Kremers, 2012; McCarthy, 1985; Ryding, 2006). All form I masdar patterns are nonconcatenative, and there is no single morpheme or feature that indicates the masdar. Form I masdar patterns include a variety of patterns: [CaCC] as in [taraka]⇒[tark] ("leave"⇒"leaving"); [CuCuuC] as in [daxl]⇒[duxuul] ("to enter⇒"entering"); [CaCiiC] as in [raħala]⇒[raħiil] ("depart"⇒"departing"); [CaCaC] as in [fariħa]⇒[faraħ] ("to be happy"⇒"being happy"); [CaCCaT] as in [kaθura]⇒[kaθraT] ("to be numerous"⇒"being numerous"); [CaCaaCaT] as in [jadura]⇒[jadaaraT] ("to be suitable"⇒"being suitable"), and [CuCuuCaT] as in [sˤaʕuba]⇒[sˤuʕuubaT].

Although there is no infinitive per se in the Arabic verbal paradigm, the base form of the Arabic verbal paradigm is generally considered to be the third-person masculine singular past tense form[7] (e.g., Ryding, 2006; Wright, 1988). For form I verbs, this has the shape [CvCvCv].[8] Verbs of this form take one of three voweling patterns [CaCaCa], [CaCiCa] and [CaCuCa]. For brevity, I will also note the meaning of the verb in infinitive form rather than the inflected form in glosses. One other issue noted on in the literature is the directionality of derivation between

7. The terms 'past' and 'present' are not used consistently in the literature for the two main active tenses in Arabic. Some grammars refer to these as 'perfective' and 'imperfective', or in some cases 'indicative' and 'subjunctive.' I will use the terms 'past' and 'present' throughout, as the aspectual nature of these two tenses is not relevant for these analyses.

8. The pattern [CaCvCa] is the underlying base pattern for all form I verbs; however, not all verbs conform to this pattern in the surface form due to regular phonological processes. The two major phonological changes are the elision of a weak consonant into a long vowel in verbs with the weak consonant in the second root position ("hollow" verbs, [CvCvCv]⇒[CvvCv]) or in the third root position ("lame" verbs, [CvCvCv]⇒[CvCvv]) (although this process is variable and some weak verbs surface as [CvCvCv]), and the gemination of the second root consonant if the verbal root is biconsonantal, [CvCvCv]⇒[CvCCv].

the verb and the masdar. Some scholars argue that the masdar is the base from which the verb is derived, in particular in light of the term 'masdar,' which means "source," and the relative unpredictability of form I verb masdars (see chapter 3 of Versteegh, 1977 for a review). However, the arguments presented in Versteegh focused on historical derivation and are somewhat unrelated to the situation of the language learner. This work, following previous traditional analyses, assumes that the 3rd-person masculine singular past tense form is the base form from which the masdar is derived.

The main question examined in this chapter is: what regularity is there in the system as a whole? For a system as a whole to be learnable, there must be some regular correspondences between verbs and their masdars. If the system is truly irregular, this presents major issues for learnability, as every masdar must be memorized individually, and masdars for a previously unseen verb cannot be predicted with any certainty. Although many analyses of the masdar system have written off the form I masdar as unpredictable, this system is not obsolete, nor does it appear to have been leveled to a single dominant pattern. However, the lack of systematic analysis of the statistical utility of the features noted in classic grammars in predicting masdar form is a major shortcoming to the claims that it is unpredictable. This chapter seeks to uncover if there are predictive features for masdar pattern for form I verbs, and what this can tell us about the learnability of the system as a whole.

In previous analyses of the Arabic masdar system for form I verbs, there are a few main features that have been indicated as related to masdar form: the phonological pattern of the verb, the transitivity and aspect of the verb, and the meaning of the verb (e.g., Ryding, 2006; Wright, 1988). In this chapter, I will examine whether these features are statistically predictive of the masdar form of the verb. First, I will examine overall statistics on the masdar patterns in the

system using a set of 2031 verb-masdar pairs. Next, I will examine whether the phonological features of the verb are predictive of masdar pattern using a comparison of predictive analogical models. Then, I will examine whether the syntactic features of the verb, namely transitivity and aspect, are predictive of masdar form. Then, I will examine whether the semantic features of the verb are predictive using a word co-occurrence model trained on Arabic sentences containing the verbs in the dataset. Finally, I will discuss the implications of these analyses for the learnability of the masdar system.

3.2 Dataset

The dataset used in these analyses was collected from the *Hans Wehr Dictionary of modern written Arabic* (Wehr, 1976), an Arabic-English dictionary that lists entries by verbal root. For each form I verb with a masdar listed, the following information was entered into a database: the past tense form of the verb, any listed alternative forms of the past tense (e.g., [ħaafa] for [ħayafa]), the verbal root consonants, the vowel pattern(s) of the present tense form, the masdar form(s) of the verb, the verb meaning(s), any prepositions used with the verb, and whether the preposition is required. The full set consists of 2764 verb listings. However, while a dictionary is a useful source for isolating verb-masdar pairs, dictionaries often contain forms that are archaic or specialized, and thus that speakers are unlikely to know. In order to create a dataset containing only masdars speakers are likely to know, the dataset was filtered to include

only masdars that occur at least once in unpointed form in the Aralex database.[9] (Boudelaa & Marslen-Wilson, 2010), a lexical database that contains word frequencies compiled using a 40 million word corpus and the Wehr dictionary. By filtering the dictionary set through a corpus, we end up with a dataset that is more representative of the language in actual use. After filtering, the dataset contains 2031 verbs.

Two corpora were also used to collate frequencies and to examine the contexts in which masdars occur. The Corpus of Contemporary Arabic (Al-Sulaiti, 2009) contains about 850,000 words, and was designed to be a balanced corpus with respect to text genre, as it contains text from a variety of print sources. A subset of Arabic Gigaword (Parker, Graff, Chen, Kong, & Maeda, 2011) was also used, which contained six months of newswire from Al-Ahram comprising 8.5 million words. Both corpora were tagged for part of speech using the Stanford log-linear Arabic POS tagger (Toutanova, Klein, Manning, & Singer, 2003; Toutanova & Manning, 2000). POS-tagging resolves some of the ambiguities introduced by undiacritized text.

3.3. Descriptive statistics on dataset

One interesting characteristic of this dataset is that a large percentage of verbs have multiple listed masdar forms. I will refer to these verbs as "multiple-listing verbs." Verbs having only one listed masdar will be referred to as "single-listing verbs." The proportion of verbs that

9. The unpointed form of the masdar was used here rather than the pointed form because Aralex has only stemmed forms available in pointed form. Specifically, *taa marbuta* is not considered part of the stem in Aralex, and thus using the pointed form conflates forms with and without *taa marbuta*, and one of the most frequent masdar classes, [CaCaaCaT], contains *taa marbuta*. Using the unpointed frequency does introduce other ambiguities (see e.g., Buckwalter, 1997). Unfortunately, either choice introduces some issues, and frequency counts from Arabic corpora must be taken as rough estimates.

have multiple masdars is quite large overall, with 488 (24.2%) of verbs in the frequency-filtered dataset having multiple listed masdars. This proportion is somewhat smaller than in the unfiltered set (35%, n=799), which suggests that some of the masdars listed in the dictionary are obsolete, or not in common use. Table 3.2 below shows the number of multiple-listing verbs with each number of masdars from this filtered set. By far, the most frequent case is verbs which have two masdars. There are also a substantial number of verbs with three masdars, but a very small number of forms with four or more masdars.

Table 3.2: Number of verbs with multiple masdars

Masdar forms per verb	**Number of verbs (% of multiple-listing verbs)**
Two	371 (76.2)
Three	96 (19.6)
Four	16 (3.3)
Five	3 (0.6)
Six	2 (0.4)

First, I will examine the distribution of masdar patterns in the single- and multiple-listing sets. The single-listing verbs show a fairly wide range of masdar patterns, with 27 in total. Figures 3.1 and 3.2 shows the type count by masdar pattern for all single-listing verbs. The 27 masdar patterns in this set show a heavy-tailed distribution. 62.6% of single-listing verbs take the most frequent pattern [CaCC] in the masdar (n=967). The second most frequent pattern is [CaCaC], with 14.5% of verbs taking it (n=225). The third most frequent pattern is [CuCuuC], with 5.8% of verbs taking it (n=90).

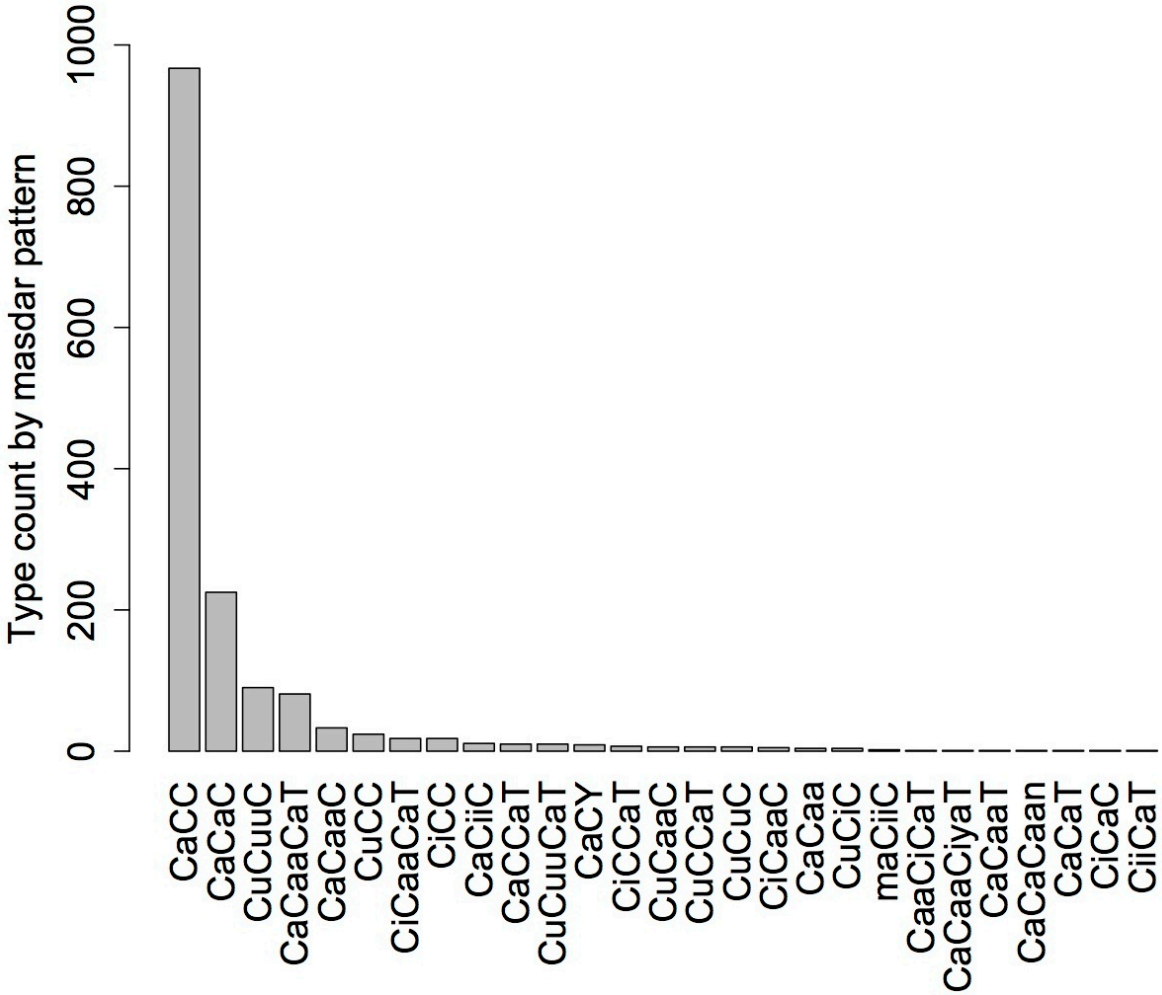

Figure 3.1: Masdar type count for all single-listing verbs

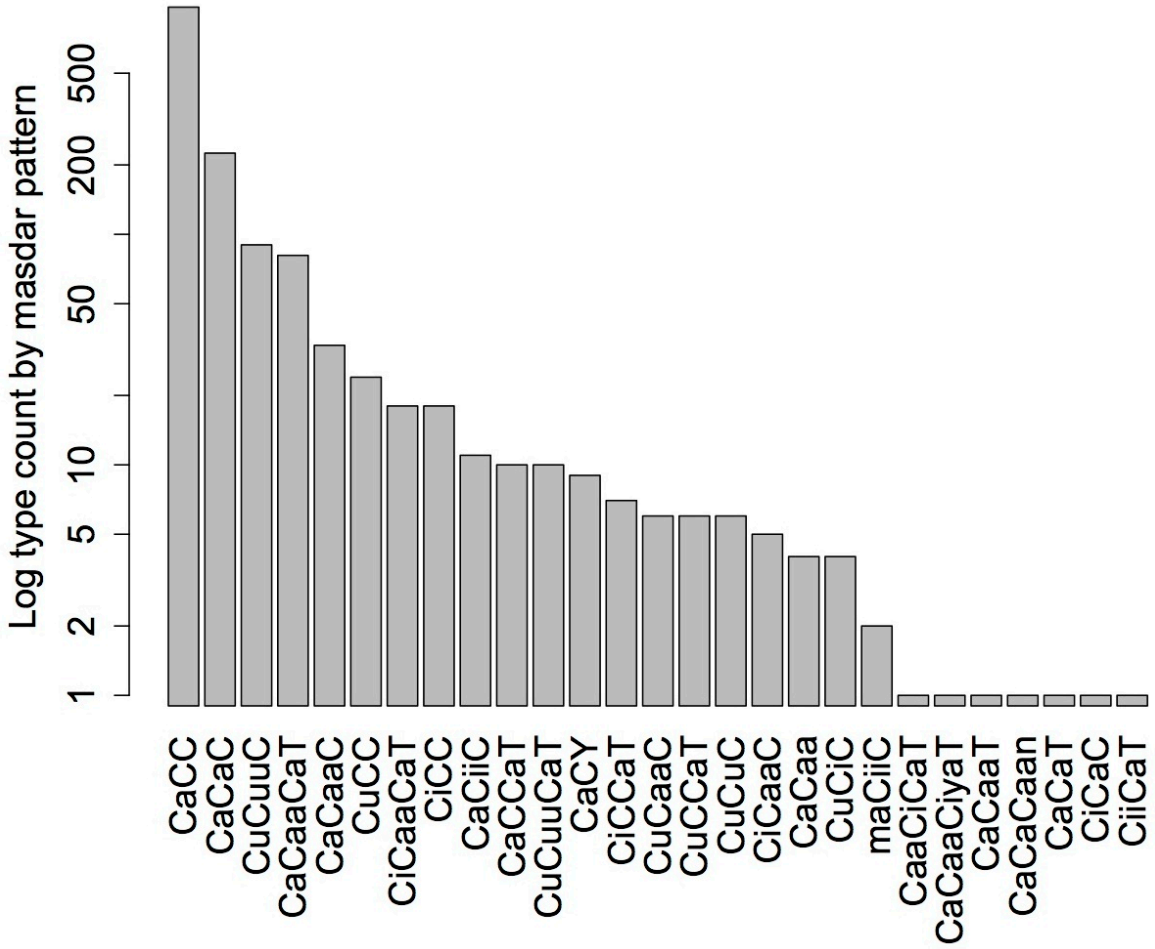

Figure 3.2: Masdar type count (log) for all single-listing verbs

The multiple-listing verbs show an even wider range of masdar patterns, with 65 in total. Many of the patterns in this dataset are not attested in Wright (1988) , and the majority of them occur very infrequently. Figures 3.3 and 3.4 shows the type count by masdar pattern for all multiple-listing verbs. Of the multiple-listing verbs, 65.6% take the most-frequent pattern [CaCC] as one of the listed masdars (n=320), but the [CaCC] masdars account for 28.5% of the masdars in this set overall, as displayed in the figure below. The second most frequent pattern is [CaCaaCaT], with 17.0% of verbs taking it as one of the listed masdars (n=83). The third most frequent pattern is [CaCaC], with 14.9% of verbs taking it at one of the listed masdars (n=73).

The multiple-listing verbs overall show a similar distribution to the single-listing verbs, with the [CaCC] the most frequent pattern by far, and [CaCaaCaT], [CaCaC], and [CuCuuC] the second through fourth most frequent patterns, though the ranking of these patterns is slightly different for the two datasets. The most obvious difference between these two datasets is the large difference in the number of patterns overall, with 27 in the single-listing verb set, and 65 in the multiple-listing verb set. However, many of the patterns in the multiple-listing verb set are extremely infrequent. If you filter this set to include only masdars that have a type count greater than five in this set, the number of patterns decreases to 28. Overall, the most-frequent masdar patterns in the single-listing verb set are also the most frequent masdar patterns in the multiple-listing set.

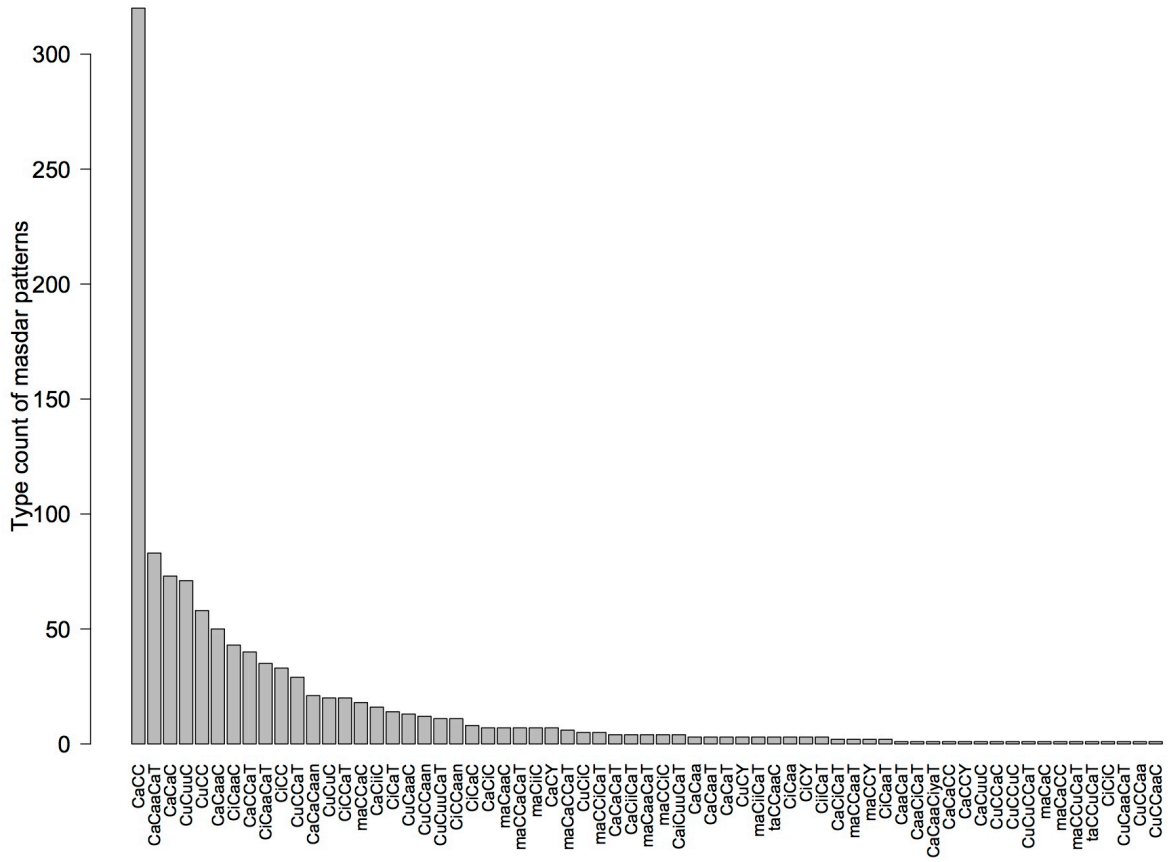

Figure 3.3: Masdar type count for all multiple-listing verbs

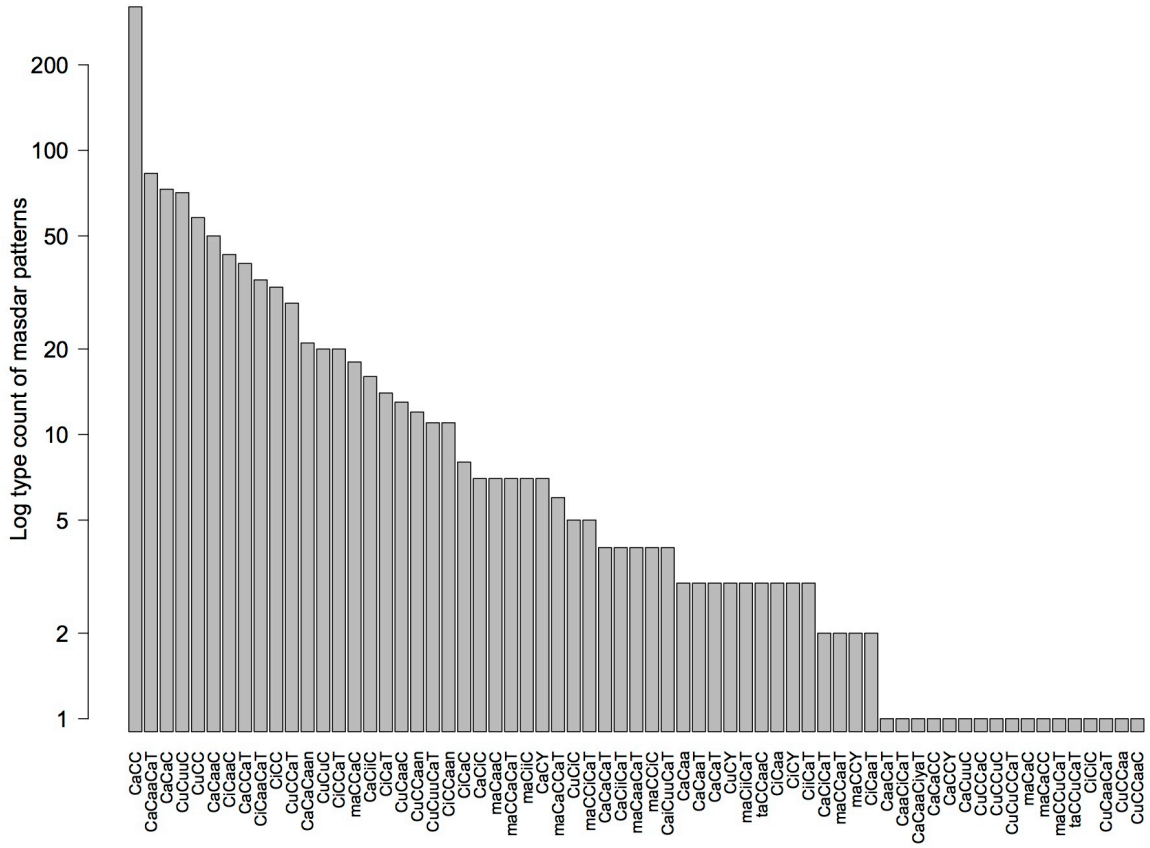

Figure 3.4: Masdar type count (log) for all multiple-listing verbs

If a verb takes multiple masdars, it is logical to expect that there is some difference between the two forms. There are a number of possible differences. For instance, Wright suggests that different masdars for one verb coincide with different meanings of the verb. It is also possible that the second masdar is the result of a form undergoing analogical leveling, for instance as in the English verb "weep", the past tense of which is somewhat instable between the historically-dominant "wept" and the regularized "weeped". A third possibility is that the different forms are dialectal variants. The first possibility of meaning differences is difficult to ascertain automatically. The second possibility can be examined to some extent through frequency of occurrence of the masdar forms. Assuming lexicographers are in touch with current usage, then the masdars should be listed in order of dominance of usage. If this is the case, then

we would expect the masdar listed first for a given verb to be more frequent than the second (or third, etc.) masdar listed.

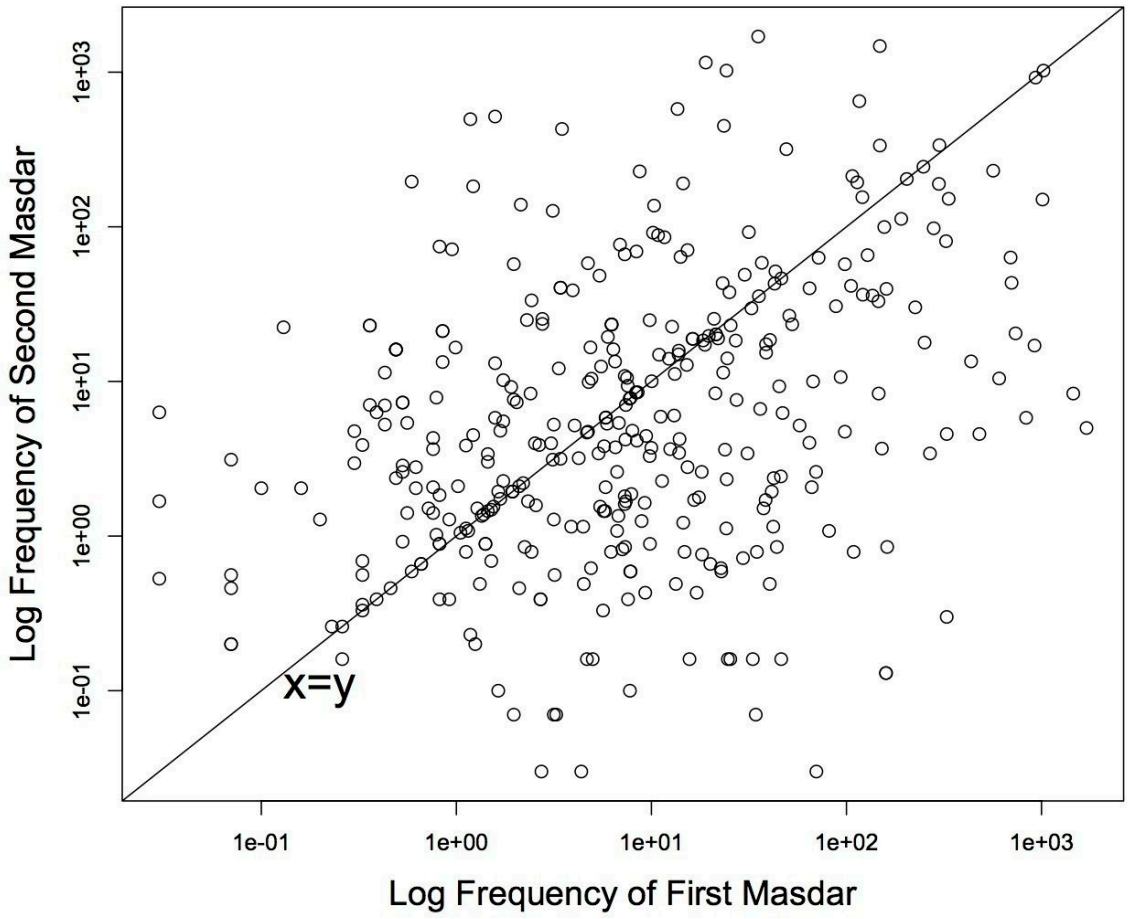

Figure 3.5: Log frequencies of first vs. second masdar for multiple-listing verbs with two masdars

Figure 3.5 shows the log frequency of the first masdar versus the log frequency of the second masdar for multiple-listing verbs with two masdars. There seems to be no consistent pattern in the frequency of the first versus second masdars, and it is not the case that second masdars are generally less frequent than first masdars, as shown in the figure above, where much of the mass occurs above the x=y line. This could happen for a variety of reasons. The most likely possibility is that masdars are not listed in the Wehr dictionary in the order of preference

or dominance, although this is common practice in dictionaries. There may be, nonetheless, dominant or preferred masdars for the multiple-listing verbs. Experiment 2 in the next chapter will examine this in more detail. For now, I will set aside the issue of multiple-listing verbs and focus on the learnability and predictability of masdars using the single-listing verb dataset only.

3.4 Phonological regularity in the masdar system

3.4.1 Descriptive statistics

As mentioned above, one major cue to masdar form indicated in the literature is the pattern of the past tense verb. As a reminder, the pattern in Arabic morphology is the CV template with the non-root consonants and vocalic melody specified. For form I verbs, there are three possible patterns, which differ in the second vowel of the vocalic melody: [CaCaCa], [CaCiCa], and [CaCuCa]. Figures 3.6 through 3.11 show the distribution of masdar patterns for each past tense verb pattern. For each verb pattern, we find a distinct distribution of masdar patterns. For [CaCaCa] verbs (n=1143), shown in Figures 3.6 and 3.7, the most frequent pattern is [CaCC], accounting for 82.3% of masdars (n=941). The second most frequent pattern for [CaCaCa] verbs is [CuCuuC], which accounts for 7.2% of masdars (n=82). For [CaCiCa] verbs (n=282), shown in Figures 3.8 and 3.9, the most frequent pattern is [CaCaC], accounting for 71.6% of masdars (n=202). The second most frequent pattern is [CaCC], accounting for 7.8% of masdars (n=22). Finally, for [CaCuCa] verbs (n=94), shown in Figures 3.10 and 3.11, the most frequent pattern is [CaCaaCaT], accounting for 63.8% of masdars (n=60). The second most frequent pattern is a tie between [CuCC] and [CuCuuCaT], which each account for 9.6% of masdars (n=9 for each).

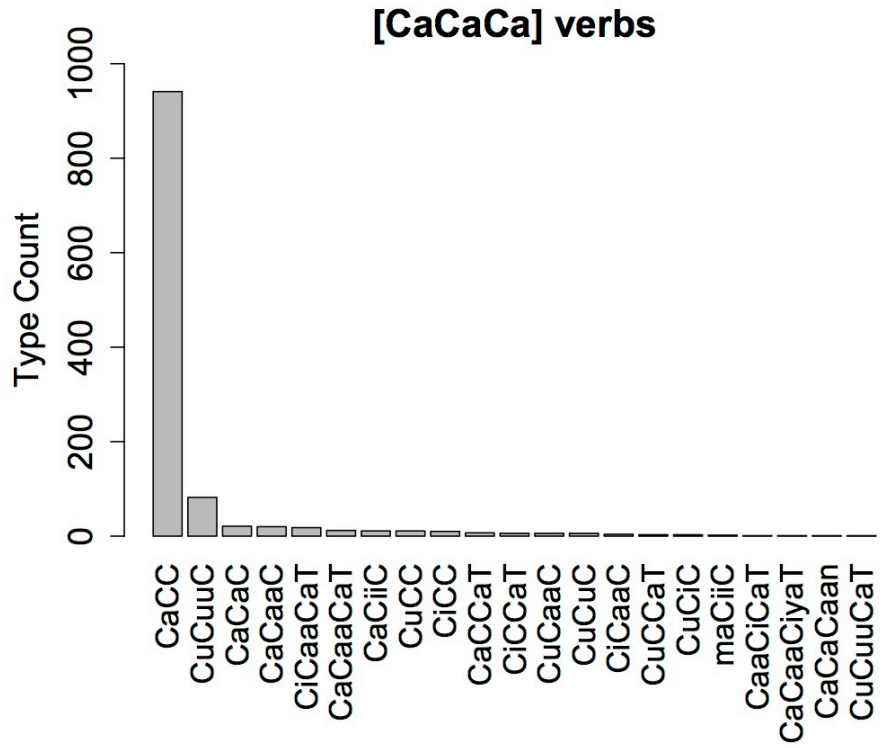

Figure 3.6: Masdar type count for single-listing [CaCaCa] verbs

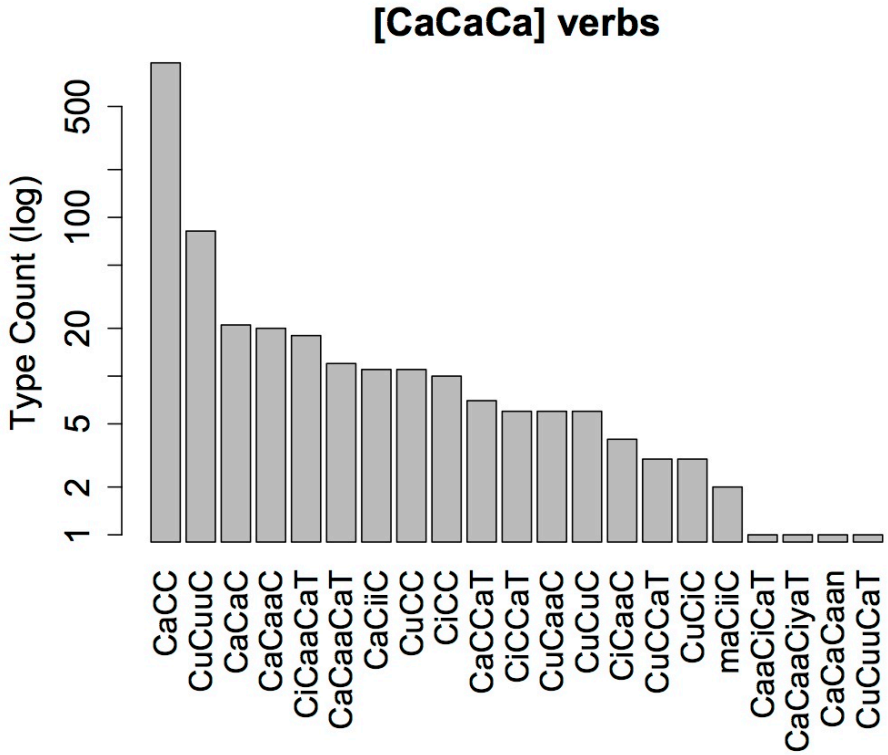

Figure 3.7: Masdar type count (log) for single-listing [CaCaCa] verbs

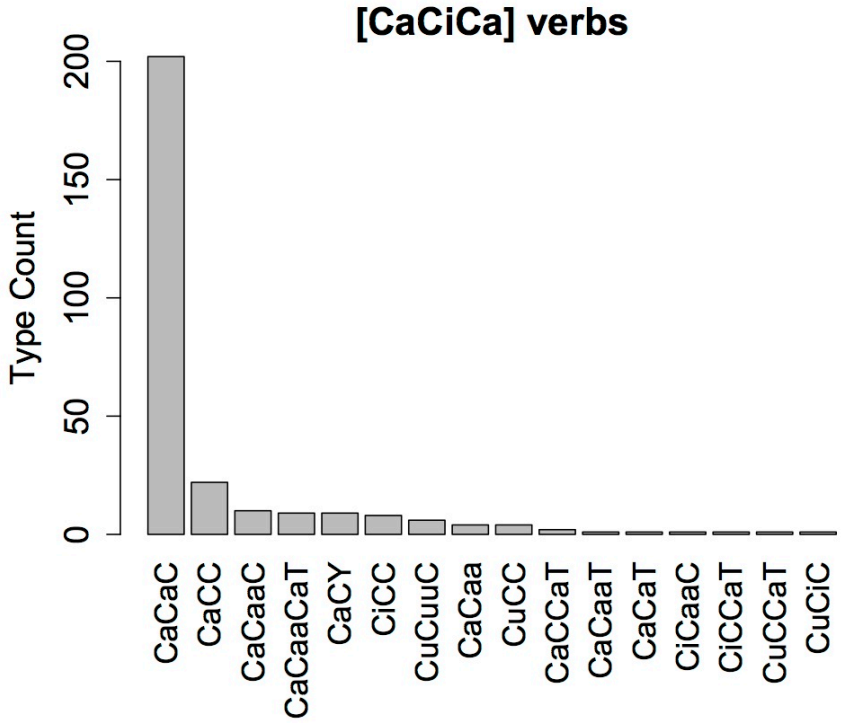

Figure 3.8: Masdar type count for single-listing [CaCiCa] verbs

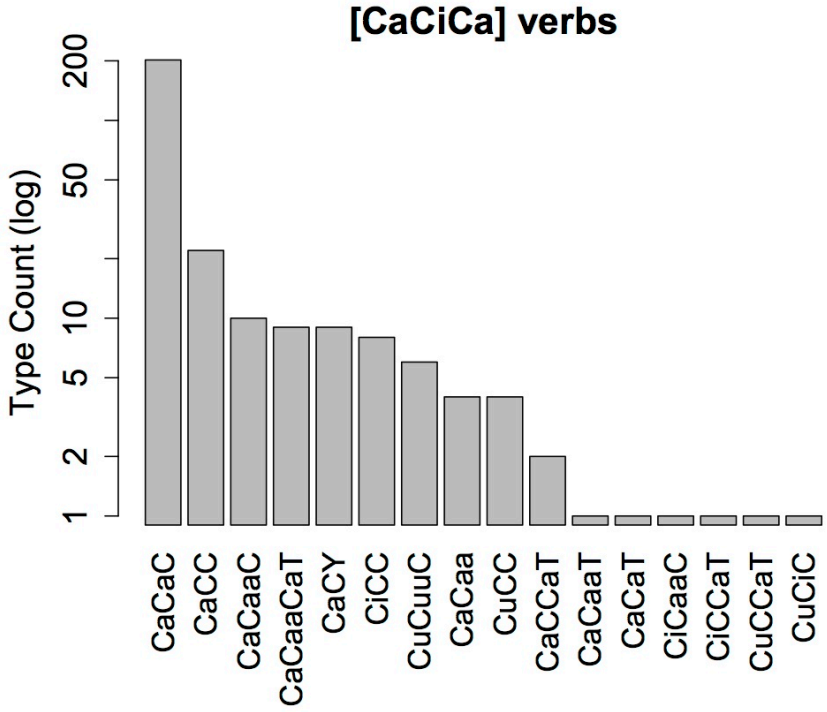

Figure 3.9: Masdar type count (log) for single-listing [CaCiCa] verbs

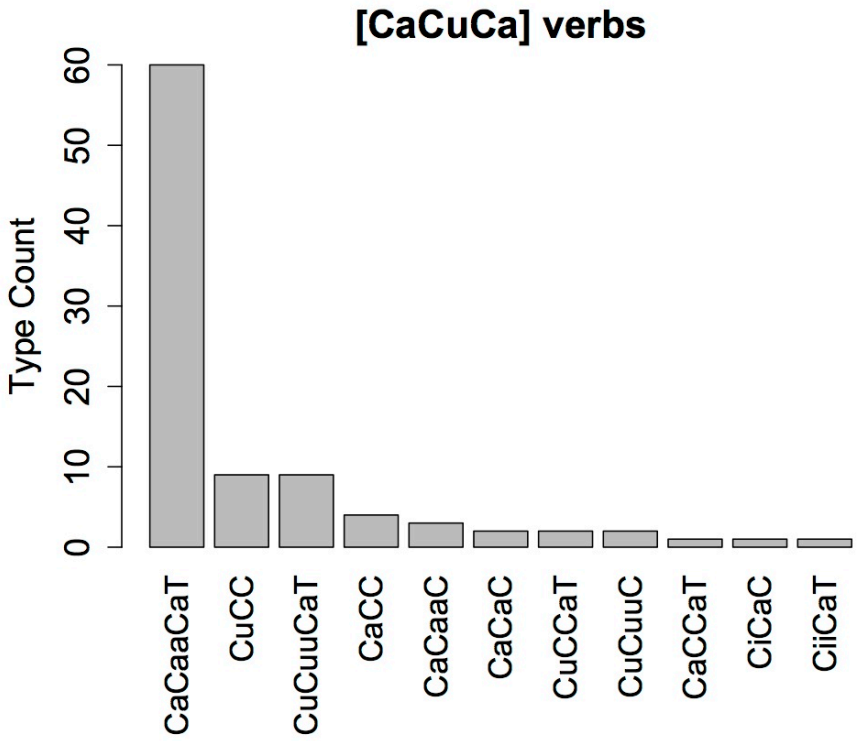

Figure 3.10: Masdar type count for single-listing [CaCuCa] verbs

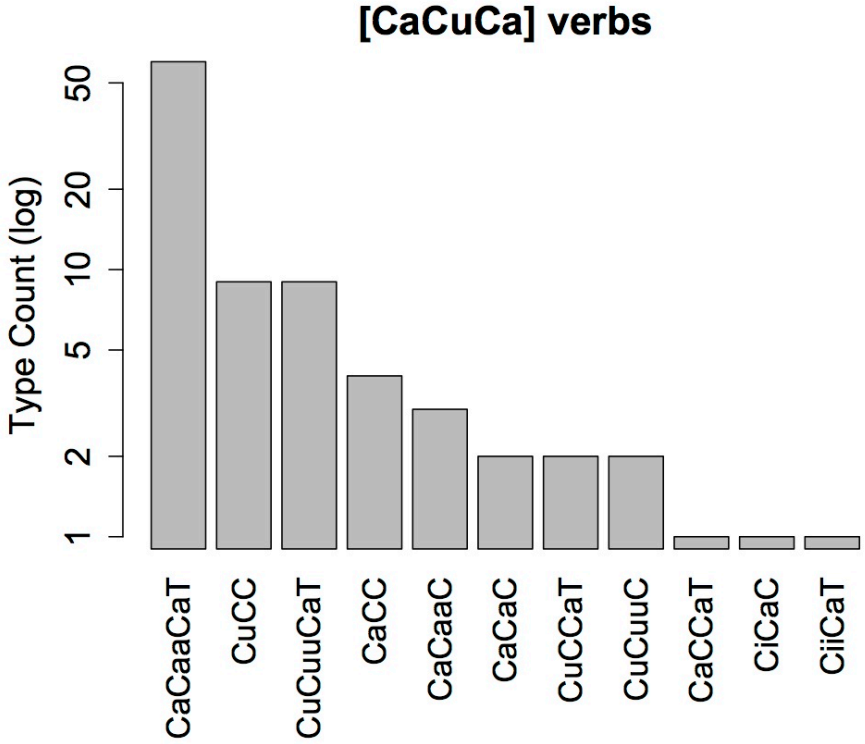

Figure 3.11: Masdar type count (log) for single-listing [CaCuCa] verbs

In total, each past tense verb pattern shows a distinct dominant masdar pattern. The dominant masdar pattern for each verb pattern, in sum, accounts for 77.9% of all masdars in the dataset. Thus, the verb pattern seems to be a fairly reliable cue to masdar pattern.

3.4.2 Analogical modeling of the masdar system

The descriptive statistics above point to a strong link between verb pattern and masdar pattern, but it is possible that there are other phonological regularities in this system that may aid in predicting masdar form. For example, there may be correspondences between the types of consonants in each verb and masdar pattern. If there are regular correspondences between the phonological features of the root consonants in the verb and the masdar pattern that verb takes, then an analogical model using fine-grained segmental features in assessing similarity between forms should perform well for this system.

To examine this possibility, I will compare three implementations of the Generalized Context Model (GCM) (Nakisa et al., 2001; Nosofsky, 1990), based on those used in Dawdy-Hesterberg & Pierrehumbert (2014). As outlined in the previous chapter, the GCM is an analogical model that predicts the best pattern to apply to an unseen form based on similarity to existing forms in the lexicon. Importantly, the predicted best pattern is a function both of similarity and of pattern strength, where pattern strength is the number of existing items taking that pattern (type frequency).

The three models are similar to the three whole gang match models from Dawdy-Hesterberg & Pierrehumbert. One critical difference between those models and the ones used here is in the definition of a gang. As noted in the introduction, I define a gang for the noun

plural as a group of forms with the same singular template that take the same plural template. However, as mentioned above, the CV template is constant for all form I verbs, while the pattern differs. Thus, the gang for these models is defined as a group of forms with the same pattern in the past tense verb that take the same pattern in the masdar. This distinction is important, as it underlines one of the important differences between the noun plural and masdar systems.

The first model is the classic *GCM*. This model predicts the masdar pattern for an unseen item among the set of all gangs in the comparison set. The unseen form is predicted to take the masdar pattern of the most similar gang in the comparison set, where similarity for each gang is a summed measure of similarity to each item in the gang, divided by similarity to all items in all gangs. In these models, the choice of gang is deterministic, as the model will always select the gang with the highest similarity measure. Thus, this model uses type statistics across the lexicon in conjunction with fine-grained segmental similarity in determining the best masdar pattern for an unseen form.

The second model, the *Pattern-restricted GCM,* is a constrained variation of the GCM. This model predicts the masdar pattern using the same metric as the GCM, but only considers gangs that have the same pattern in the past tense verb as the test item. Among those gangs, it selects the gang with the highest similarity measure. Thus, this model uses type statistics on the verb pattern in conjunction with fine-grained segmental similarity in determining the best masdar pattern for an unseen form.

The third model, the *Simple Pattern Match*, predicts the masdar pattern using type statistics on the verb pattern, without considering additional fine-grained segmental similarity. It simply predicts that an unseen form will take the pattern of the largest gang with the same verb pattern.

Ten-fold cross validation was used to ensure that model performance is stable across different test sets (Breiman, Friedman, Olshen, & Stone, 1984; Mosteller & Tukey, 1968; Stone, 1974). Under this protocol, a randomly-selected 25% of the dataset was selected as the test set, and the remaining 75% was used as the comparison set. The model makes a prediction for the masdar form of each verb in the test set based on similarity to the verbs in the comparison set. Since these were all real words, and the masdar is known, accuracy is calculated as the number of times the model predicts the correct masdar pattern for a test verb. These models are deterministic, and always select the most-likely masdar pattern for a test verb, as this results in the highest likelihood of accuracy. This procedure was iterated ten times, each time with a random 25% test set, and performance was averaged across the ten rounds.

The dataset consists of 1408 verb-masdar pairs. The pairs were classified by gang, where a gang is a set of forms with the same pattern in the past tense and the same pattern in the masdar. One example of a gang is [CaCaCa]⇒[CaCiiC], which contained pairs such as [raħala]⇒[raħiil] and [ʃaxara]⇒[ʃaxiir]. The dataset used here is slightly smaller than the one above, as the cross-validation procedure requires that each gang contain at least 4 forms. Thus, items from some rarer gangs were not included in this set. The set contained in total 23 gangs, with the largest gang containing 949 items, and the smallest containing 4.

A random baseline was also established to assess whether models overall performed above chance. The baseline model predicts the masdar for a test item by selecting a random gang, weighted by gang size.

Figure 3.12 shows the model performance across ten rounds, with error bars indicating the S.E. Baseline performance is indicated by the solid line, and baseline S.E. is indicated by the dashed lines. First, the *GCM* performs significantly worse than the two pattern-based models, the

Pattern-Restricted GCM and the *Simple Pattern Match*, t(9.64)=-27.27, p<0.001. Second, the *Pattern-Restricted GCM* does not perform significantly better than the *Simple Pattern Match*, t(9)=-0.94, p=0.37. All three models perform well above the baseline of 47.3% accuracy.

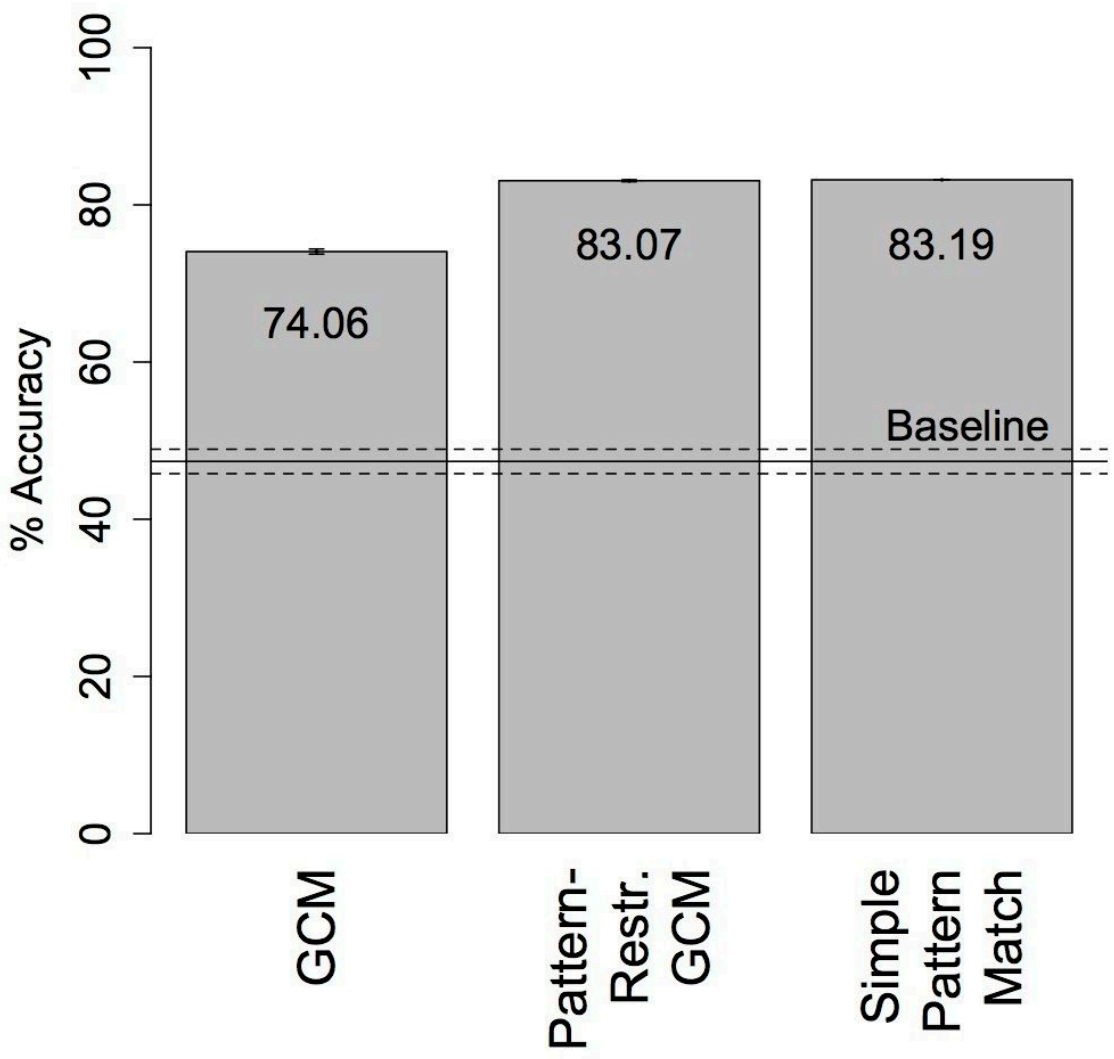

Figure 3.12: Accuracy for all models on masdars

These results indicate two things. First, they indicate that types statistics on the verb pattern are a very reliable cue to masdar form, which was predicted from the overall distribution

of masdar patterns by verb pattern in the previous section. Second, these results suggest that there are not reliable phonological cues to masdar form beyond those cues provided by the verb pattern. Similar modeling work on the noun plural system found that the template-restricted GCM, which uses fine-grained segmental similarity, was better at predicting the plural for an unseen singular than a model which only used type statistics on the singular template; however, the effect size of segmental similarity in predicting the noun plural was very small, adding only about 2% accuracy (Dawdy-Hesterberg & Pierrehumbert, 2014). Thus, it is possible that the set of verb-masdar pairs used in this analysis was too small to pick up on effects of fine-grained segmental similarity, which may be very weak to begin with. Nonetheless, this indicates that the primary reliable phonological cue to masdar form is the verb pattern, and that the specific quality of the verbal root consonants does not seem to be informative in this analysis. This finding is in line with Wright's (1988) analysis, which mentions only verb pattern and no additional phonological factors. In the next section, I will examine whether the syntactic and semantic factors Wright indicates also play a role in masdar formation.

3.5 Syntactic regularity in the masdar system

I demonstrated in the previous section that the verb pattern is a major predictor of masdar form, and that a model using type statistics on this factor achieves 83.2% accuracy in predicting the masdar for existing verbs. As noted in the introduction, the classic Arabic grammars mention a few other factors that may be predictive of masdar form. Specifically, two others factors pointed to in Wright are syntactic and semantic features of the verb.

The syntactic features mentioned by Wright are the transitivity and aspect of the verb, and are referenced in conjunction with the verb pattern. According to Wright, transitivity is linked to four masdar patterns: transitive [CaCiCa] and [CaCaCa] verbs take [CaCC]; intransitive [CaCaCa] verbs take [CuCuuC]; and intransitive [CaCiCa] verbs take [CaCaC]. Aspect governs two patterns: stative [CaCuCa] verbs take [CaCaaCaT] and [CuCuuCaT]. However, it is noted in the grammar that vowel pattern is not independent of verb transitivity and aspect, with most [CaCaCa] verbs being transitive, most [CaCiCa] verbs being intransitive, and most [CaCuCa] verbs being intransitive and stative.

In order to examine the extent to which verb pattern is linked to transitivity and aspect as well as whether these influenced masdar pattern, I used the AraComLex database (Attia, Pecina, Toral, Tounsi, & van Genabith, 2011) in conjunction with my masdar database. The AraComLex database provides information on verb transitivity for 802 of the 1543 verbs in the dataset. The verbs in AraComLex were classified for transitivity using a Multilayer Perceptron model (Haykin, 1998; Rosenblatt, 1961) trained on a set of manually annotated verbs. The model achieved high precision and recall on classifying verbs for transitivity (0.85 for both), so this is a fairly reliable source of information on verb transitivity. To my knowledge, there is no other database that contains information on Arabic verb transitivity, so this database proved very fruitful for these analyses. I marked the verbs for aspect, specifically stativeness, using a simple measure: whether the verb meaning in the database contained "to be _____." Thus, verbs such as [rasaxa] "to be firmly established" and [makana] "to be or become strong" were marked as stative, and all others not containing the phrase in question were marked as non-stative.

First, with regards to transitivity, shown in Table 3.3, we see a clear split among the verb patterns. The majority of [CaCaCa] verbs are transitive, with nearly 80% of verbs in the set being

transitive. For [CaCiCa], there is a slight tendency toward intransitivity. For the [CaCuCa] verbs, there is a nearly even split between transitive and intransitive verbs, but there are very few verbs of this pattern in this dataset. Note that the statistics in Table 3.3 were drawn from the subset of verbs for which there was transitivity information in AraComLex, while the statistics on aspect in Table 3.4 were drawn from the entire dataset, which is substantially larger.

Table 3.3: Transitivity of verb patterns

Verb Pattern	Transitive	Intransitive
CaCaCa	476 (78.8)	128 (21.2)
CaCiCa	71 (42.0)	98 (58.0)
CaCuCa	14 (48.3)	15 (51.7)

Second, for aspect, shown in Table 3.4, there is also a clear split among the verb patterns. The majority of [CaCaCa] verbs are non-stative, while the majority of [CaCuCa] verbs are stative. [CaCiCa] verbs show a slight tendency to be stative, but not overwhelmingly so. If we combine the information in these two tables, we find that transitivity and aspect are very strongly linked to the [CaCaCa] and [CaCuCa] verb patterns, but that the [CaCiCa] verbs are not strongly defined by either of these factors, with a slight tendency toward intransitivity and stativeness.

Table 3.4: Aspect of the verb patterns

Verb Pattern	Stative	Non-stative
CaCaCa	137 (11.7)	1030 (88.3)
CaCiCa	161 (57.1)	121 (42.9)
CaCuCa	84 (89.4)	10 (10.6)

For the [CaCaCa] and [CaCuCa] verbs, there are strong associations between the verb pattern and transitivity and aspect, respectively. Given that the verb pattern is also strongly associated with the masdar pattern, with the majority of [CaCaCa] verbs taking [CaCC] masdars and the majority of [CaCuCa] verbs taking [CaCaaCaT] masdars, this is evidence in support of Wright's assertion that transitivity and aspect are related to masdar formation. However, a learner does not need to know the transitivity or aspect of a particular verb in order to learn the masdar because of this association between these features and the two verb patterns.

However, the fact that the [CaCiCa] verbs are not strongly associated with either transitivity or aspect warrants further investigation. Although it is the case that most [CaCiCa] verbs take the masdar pattern [CaCaC], Wright posits that transitive and intransitive verbs of this pattern take distinct masdar patterns. Figure 3.13 shows the proportion of intransitive and transitive [CaCiCa] verbs taking each masdar pattern. As can be seen in this figure, Wright's assertion that intransitive [CaCiCa] verbs primarily take [CaCaC] and transitive [CaCiCa] verbs primarily take [CaCC] is not accurate; in fact, the distribution of masdar patterns is extremely similar for both transitive and intransitive verbs. Likewise, as shown in Figure 3.14, aspect does not have a clear effect on the masdar form of [CaCiCa] verbs. Thus, it does not appear that transitivity or aspect has an independent effect on masdar pattern, as Wright suggested. Rather, there is a strong association between the verb patterns and the transitivity and aspect of the verb, and the verb patterns in turn have a strong association with the masdar patterns. Overall, the syntactic features of the verb do not appear to add predictive power beyond that given by the verb pattern alone.

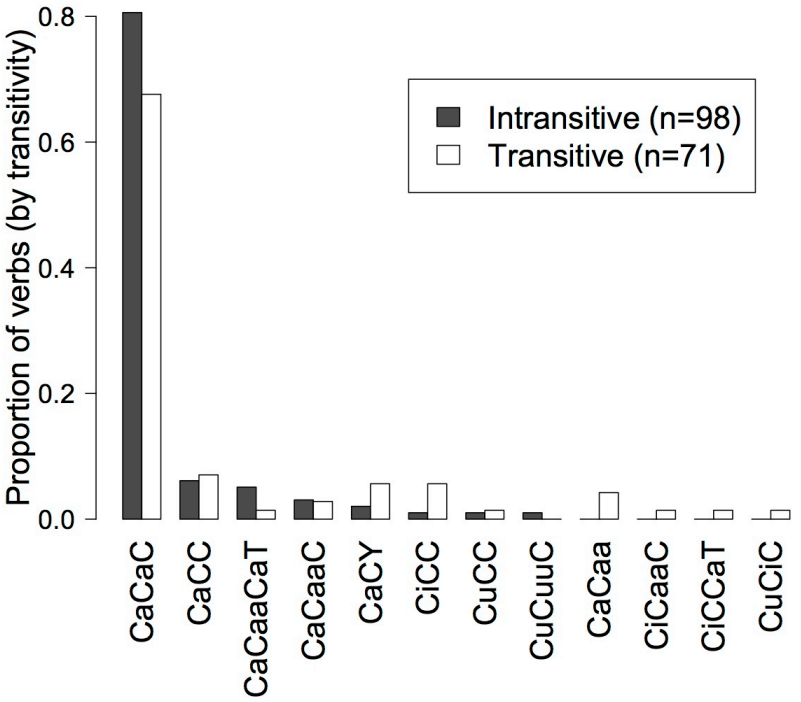

Figure 3.13: Proportion of intransitive and transitive [CaCiCa] verbs by masdar pattern

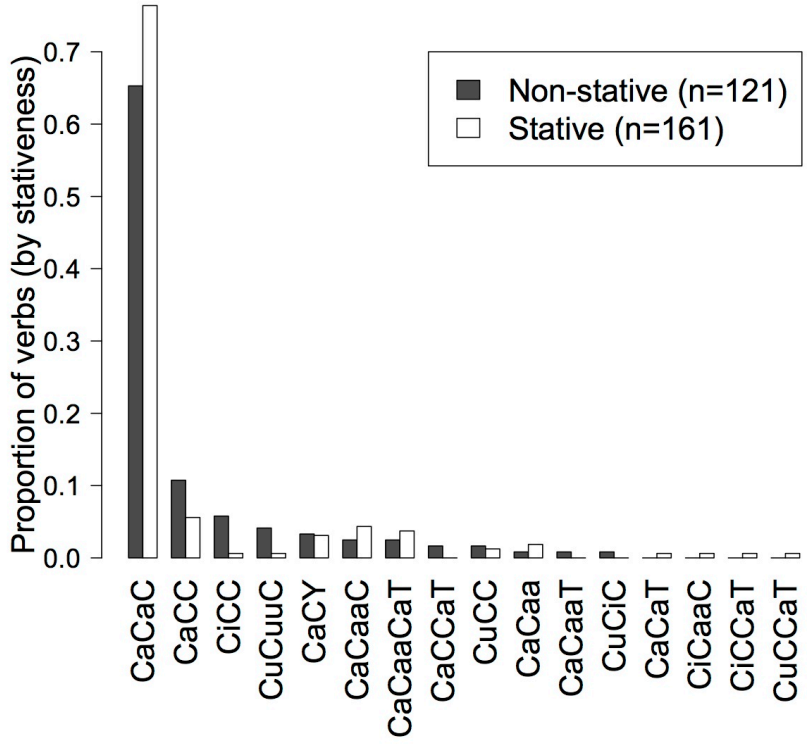

Figure 3.14: Proportion of non-stative and stative [CaCiCa] verbs by masdar pattern

3.6 Semantic regularity in the masdar system

The second factor mentioned in Wright is the semantic features of the verb. What is meant by 'semantic features' is extremely variable, and seems to be a small number of groups of verbs with particular meaning associations. In addition to the classes dictated by verb pattern and syntactic features, Wright lists an additional six masdar classes that are associated with specific classes of meaning: [CiCaaC] for flight and refusal, [CuCaaC] for sickness or ailment, [CaCaCaan] for violent or continuous motion, [CaCiiC] for change of place, [CuCaaC] and [CaCiiC] for sound, and [CiCaaCaT] for office, trade or handicraft. While there are examples from each class that fall into the scope of meaning listed, there are also examples that do not conform to the meaning, or conform to the meaning but take a different masdar pattern (see also Grenat, 1996). It is difficult to find a unified approach to the semantics with classes as disparate in meaning as these, in particular when other classes do not seem to involve the semantics at all. This section examines the extent to which these semantic features of the verb noted by Wright are predictive of masdar form.

There are many automatic approaches to extracting meaning representations from raw text. Many of these approaches stem from the general observation that similar words appear in similar contexts (e.g., Harris, 1954). Thus, words that appear in similar contexts are likely to be similar in meaning. There are a number of machine-learning approaches to semantic representation estimation which function on a variety of different architectures. For instance, Latent Semantic Analysis (LSA) and Hidden Markov Models (HMMs), and Neural Network Language Models (NNLMs) all capitalize on this observation. Generally, these approaches use the contexts (surrounding words) in which a word appears to generate mathematical vectors that roughly correspond to the meaning representation of the word, such that a given word will be

closer in the *n*-dimensional vector space to semantically related words than to semantically unrelated words. These word vectors can also be used to group words into semantic classes, or to disambiguate polysemous or homographic words. There are many different algorithms that use word co-occurrences to represent semantics, and they differ in many cases in terms of the size of the context considered. If one wishes to capture thematic relations between words, a larger context may be useful, as thematically-related words may not appear very close to the target word, and may even appear in different sentences. If one wishes to capture more local relationships among words, for instance verb argument structure, then an approach which focuses on very nearby words to the target word would be more appropriate.

There has been some success in using word co-occurrence models to predict morphological variants in English. Recent work by Mikolov and colleagues used a novel NNLM-based approach to predict a variety of morphosyntactic variants of English words, including predicting the gerund from the verb, using learned word pair relationships (Mikolov, Chen, Corrado, & Dean, 2013; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). For example, given the word vectors for [think] and [thinking] (and similar pairs), the model can predict, e.g., [reading] from [read] with about 60% accuracy[10]. Although the gerund system in English is extremely regular, unlike the Arabic masdar system under examination, this approach does approximate what I wish to achieve in terms of predicting a morphological variant of a given word using known word pair relationships.

In Arabic specifically, there has been some previous success in using word co-occurrence methods to classify verbs into semantic classes. Snider and Diab (2006) combined word vectors

[10] Note that the accuracy is aggregated across a variety of different morphosyntactic relational types, of which the gerund is only one of nine. The authors do not break down accuracy for each specific relational type, so this accuracy can be taken only as a rough estimate.

from an LSA model trained on a text corpus and verb argument structures from an Arabic Treebank to predict verb semantic clusters. The authors found that this approach yielded clusters that were significantly closer to the gold standard verb semantics clusters than a random baseline[11]. Although the aims of this work were different than the aims of the current analyses, the success of Snider and Diab's approach validates using word vector approaches to capture Arabic verb semantics.

In order to ascertain whether the semantics of the verb play a role in masdar form, I adopt a similar approach. However, rather than clustering verbs by their semantics as above, the analyses I detail below attempt to predict the masdar for a given verb using the word vectors of the verb generated by a word co-occurrence model trained on sentences containing the verbs. If semantics are relevant to masdar formation, then verbs with similar semantics according to this approach should also take similar masdars. These analyses feed the word vectors given by the trained model into several classification algorithms to ascertain whether these semantic features are predictive of masdar form, beyond the predictiveness of the verb pattern demonstrated in the previous section.

Hidden Markov Models have shown success in semantic feature extraction for a variety of language modeling purposes, including semantic disambiguation (Huang et al., 2014). For these analyses, I use a 2nd-order HMM. HMMs, generally, are statistical models that model a system of hidden states that correspond to observed sequences (Rabiner, 1989). For our

[11] The authors report F_β, which is a measure of the overlap between the model-generated clusters and the gold standard clusters weighted equally by precision and recall, rather than accuracy, so it is difficult to compare these results directly to those of similar models. The critical point is that Snider and Diab's approach did yield positive results in classifying Arabic verbs by semantic features, which suggests that a similar approach should be able to aid in predicting masdars for form I verbs if the semantics are involved in masdar formation.

purposes, given a sentence, the model attempts to find the transitional probabilities between hidden states that correspond to the observed sequences, where each observation is a word in the sentence. The model is trained on multiple instances of each verb, such that the output is an approximation of the general word co-occurrences around the verb. The output from the trained model is a vector of 25 numbers for each input word in the sentences. Thus, for each verb, we have a vector of 25 numbers that, in aggregate, represent the transitional probabilities between the preceding word and the target verb, and between the target verb and the following word. Although the vectors themselves do not correspond to specific words and/or states, as noted above, verbs that occur in similar contexts should be closer in the vector space than verbs that occur in dissimilar contexts. By extension, then, semantically-related verbs should be closer in the vector space than semantically-unrelated verbs. This approach, which uses a narrow context window relative to that of Snider and Diab or Mikolov et al., was intended to primarily focus on lexical semantics, rather than general topic.

The HMM was trained on sentences from Arabic Gigaword containing the verbs under investigation. First, frequency counts for all verbs in the masdar database were collected from the POS-tagged Gigaword. The verb frequency counts were collapsed across the verbal paradigm, with all persons, numbers and genders included, but not across verb tenses. Only tokens in the past tense were counted and analyzed. Next, sentences containing verbs with a token frequency of ≥5 in the past tense were extracted for analysis. Prefixes and suffixes that were not part of the conjugated verbs were separated from the verb by a space during this process. For example, direct object pronouns were separated using this approach, as were prepositions that attach to the verb as prefixes. Finally, affixes that specified person, number, and gender were deleted, such that the remaining verb was in the 3rd person masculine form.

This procedure resulted in 96,971 sentences, which contained 374 of the verbs from the original set. The verbs took 17 different masdar patterns, with 64.4% (n=241) taking CaCC, 9.4% (n=35) taking CaCaC, and 3.2% (n=12) taking CaCaaCaT. Thus, the distribution of masdars was similar to that in the full dataset. The HMM was trained using code from Yang, Yates & Downey (2013) and the entire set was trained on a single node.

The output vectors for the target verbs from the trained HMM were analyzed using several classification algorithms. The masdar pattern of the verb was supplied to the model as the desired output, but was not available to the model except for verification of accuracy. Thus, the classification algorithms had access to 26 factors. Accuracy for each algorithm is how many masdars could be correctly predicted by the algorithm for the verbs in the set. In addition to the vectors from the trained HMM, the classification algorithms were also supplied with the pattern of the verb. Based on the proportion of the majority class [CaCC] in the dataset, if a model always predicted the majority class, then it would achieve 64.4% accuracy. If a model used only the vowel pattern as a factor, then it would achieve 77.0% accuracy. The main factor of interest is whether the semantic vectors add additional predictiveness for the masdar classes other than [CaCC], [CaCaC], and [CaCaaCaT], and in particular for those classes mentioned by Wright as having specific meaning associations.

A variety of algorithms were tested on the dataset from the available classification algorithms in Weka (Hall et al., 2009), an open-source data-mining program. All algorithms were tested using 10-fold cross validation. The best-performing model was the Attribute Selected Classifier using the Naive Bayes Simple algorithm (Rennie, Shih, Teevan, & Karger, 2003), which achieved 73.8% accuracy in classifying the verbs by masdar. Of note is that this model only gave three classifications total, although there were 17 masdar patterns in the dataset. The

model classified all verbs as taking one of the three dominant patterns by vowel pattern, [CaCC], [CaCaC], and [CaCaaCaT]. Importantly, this model in fact ignored all of the semantic vectors, and used only the vowel pattern in assigning predicted masdar patterns, which indicates that the semantic features provided no predictive power under this approach.

Clustering approaches were also examined, as they are often used on semantic vectors like those in this dataset to find semantic groupings automatically. The best-performing clustering approach was achieved with the Classification via Clustering method using the Simple *k*-Means algorithm (MacQueen, 1967). In general, *k*-means clustering is a minimization function that attempts to find the set of clusters wherein the mean distance between each data point assigned to a cluster and the centroid of that cluster is the lowest possible across all clusters, given a pre-specified number of clusters. The number of clusters was varied, and a model was built for each number from 2 to 17. When 17 clusters was specified, the model correctly classified only 25.9% of verbs. The best performance for this model was achieved when the number of clusters was 2, achieving 51.6% accuracy. When the number of clusters was set to 2, the model achieved good success on the [CaCC] (72.2% accuracy) class, and limited success on the [CaCaC] (40.0% accuracy) and [CuCuuC] (13.2% accuracy) classes, but did not correctly classify instances of any other class. Overall, the classification approaches that used the semantic vectors from the HMM in predicting masdar form significantly worse than the approach which ignores the semantic vectors entirely, and in fact perform significantly worse than a simple baseline of always choosing the most frequent pattern.

In sum, using an HMM trained on a large set of sentences from Arabic Gigaword, semantic features that might be relevant in predicting masdar form (if they truly exist) cannot be identified. The best-performing classification algorithm, in fact, did not use the word vectors

from the trained HMM in classifying verbs into masdar classes, but rather relied only on the vowel pattern. This supports the conclusion from the previous section that the verb pattern is the primary factor in predicting masdar form. It is possible that the HMM-based approach did not appropriately capture the relevant semantics of the verb; however, given the fact that the HMM output vectors were actually detrimental to masdar prediction, and that previous literature (e.g., Grenat, 1996) has indicated skepticism about the semantic classes Wright noted, no further analyses on the semantics were performed. The third chapter, which examines generalization of masdar patterns to nonce forms, may provide more insight into whether the semantics play any role in masdar formation, but given the analyses here, the evidence is leaning away from the semantics being predictive of masdar form.

3.7 Discussion

In this chapter, I have demonstrated that the Arabic masdar system is not unpredictable, as has been previously claimed. The phonological representation of the verb pattern plays the primary role in masdar formation, with each of the three verb patterns showing distinct distributions of masdar patterns, and each having a different dominant masdar pattern. An analogical model that uses type statistics on the verb pattern predicts the masdar pattern for unseen verbs with about 80% accuracy. I have also shown that the syntactic features of the verb noted by Wright do not have an independent effect on masdar form. Analyses of the transitivity and aspect of the verbs in the dataset show that there is a strong association between the verb patterns and transitivity and aspect, but that these syntactic features do not predict masdar form beyond what is predicted by the verb pattern alone. In addition, using machine-learning

approaches to automatic semantic classification, I have shown that there is little evidence for the influence of semantics on masdar form as outlined by Wright (1988). In fact, classification algorithms using the semantic features from the trained HMM are worse at predicting the masdar form for the verb than algorithms which attend only to the phonological feature of the verb pattern. Although these analyses are not definitive, they strongly suggest that the semantic features of the verb are not predictive of masdar form.

These analyses show that the system is learnable to a large extent using only the type statistics on the verb pattern. The best-performing analogical model of the masdar system, which achieved 80% accuracy in predicting masdars for unseen verbs, in fact outperforms the best-performing analogical model of the noun plural system from Dawdy-Hesterberg and Pierrehumbert (2014), which achieved 65% accuracy in predicting plurals for unseen singulars. As demonstrated in the previous chapter, the noun plural system is very learnable, with speakers achieving high accuracy on filler items in the nonce-form task, as well as using a variety of different plural patterns productively in generalizing to new forms. Thus, the masdar system should be quite learnable. Moreover, speakers should generalize masdar patterns to new forms in a manner that reflects knowledge of these type statistics on the verb pattern. In addition, the factors that are predictive of masdar form are arguably simpler than those predictive of noun plural form. For the masdar, type statistics on the abstract phonological representation of the verb pattern are the only demonstrable predictor of masdar form, whereas for the noun plural, fine-grained segmental features in conjunction with type statistics on the abstract CV template of the singular noun are the most predictive of the plural. This factor suggests that the masdar may be, in fact, more easily learned than the noun plural system, given equal exposure on the part of learners.

A remaining question is why so many verbs appear to have multiple existing masdars. One possibility, raised by Wright, is that different masdars of a verb correspond to different meanings. A second possibility is that both forms have the same meaning, but are dialectally variant. A third possibility is that one of the forms is obsolete or undergoing leveling, and one of the masdar forms is dominant or preferred.

The next chapter will examine these predictions experimentally. First, I will examine the question of verbs with multiple masdars with an experiment using a forced-choice paradigm on verbs with two masdars. Second, I will examine the learnability and generalizability of existing masdar patterns using a nonce-form task. In conjunction with the analyses presented in this chapter, the experiments in the next chapter will illuminate the extent to which the Arabic masdar system is truly learnable.

Chapter 4 : Learnability and generalization of Arabic masdars

4.1 Introduction

The Arabic masdar system, despite previous claims in the literature, is quite predictable on the basis of type statistics on the verb pattern. As demonstrated in the previous chapter, the masdar can be predicted for about 83% of verbs using an analogical model that selects the most frequent masdar pattern among verbs in the lexicon sharing the same pattern. Analyses of the transitivity and aspect of the verb revealed no independent predictiveness of masdar form beyond that defined by the verb pattern. In addition, using an HMM trained on a large corpus set, I showed that there appears to be no additional predictability from the semantics of the verb. Thus, the masdar for a given verb is, in fact, quite predictable on the basis of a coarse-grained phonological representation in conjunction with the type statistics for existing verb-masdar pairs in the lexicon.

However, the previous chapter only examined predictability for existing verbs, and the extent to which specific cues are available to speakers in determining masdar form. Speakers do not always select deterministically among the choices available to them as the analogical models in the previous chapter do, nor do they necessarily use all of the predictive cues available to them (as in experiment 1A with the noun plurals). Further, the masdar system presents another challenge which was not fully addressed in the previous chapter: verbs which have multiple masdar forms. In this chapter, I will examine two main questions. First, I will examine whether the multiple-listing verbs truly have multiple masdars, and if so, what is the source of the variability. Second, I will examine the extent to which native speakers know, and can generalize, the existing masdar patterns in the language to new forms.

In experiment 2, I will examine speaker preference for the masdars of multiple-listing verbs. The first question this experiment seeks to answer is whether both masdars for multiple-listing verbs with two masdars are truly active in the lexicon. If both forms are active in the language, then a second question arises, which is what the basis of this split is. As mentioned in the previous chapter, there are a number of possible reasons why a verb may have multiple masdars that appear in a corpus. First, this may be a result of a dialect difference in which masdar is used for a given verb, where some dialects use one form and some dialects use another. Second, this may be the result of analogical leveling, where one form is being replaced with another, but the leveling is incomplete. Third, it may be the case that one of the forms is in fact obsolete, but still appears in the corpus because the corpus contains text from older sources. One other possibility is that there may be some difference in semantics between the two masdars, although this seems unlikely given the demonstrated lack of semantic effects on masdar form in the previous chapter. In order to disentangle whether these verbs in Arabic truly have multiple active masdars, participants in this experiment will be asked to select the preferred masdar for multiple-listing verbs that have two masdars. If the existence of two masdars for a given verb is a result of lexical differences between dialects, then speakers from different dialect backgrounds should prefer different masdars. If it is a result of analogical leveling, then speakers should generally prefer the dominant form, although they may not show complete preference; for instance they may prefer it only 75% of the time**.** If it is a result of a completed leveling, then speakers should always prefer the dominant form. This experiment will help to disentangle the extent to which multiple masdars for a given verb are active in Arabic, and if so, why this is the case.

Experiment 3 will examine generalization of the existing masdar patterns in Arabic to unseen items in a nonce-form experiment. Like the noun plural experiments, this experiment will examine two facets of generalization: which linguistic factors speakers attend to in determining the possible masdar patterns for an unseen verb, and how speakers select among the possible patterns they have determined. The masdar system is, in fact, more predictable than the noun plural system, so this provides an interesting counterpoint to experiments 1A and 1B, in which speakers chose probabilistically among the possible plurals. If the system is more predictable, then speakers are more likely to choose deterministically among the possible choices (e.g., Culbertson et al., 2012; Schumacher et al., 2014). Second, this experiment will examine which linguistic factors speakers utilize in determining the best masdar pattern to apply to an unseen verb. Although the model comparison showed that type statistics on the verb pattern are strongly predictive of masdar pattern for existing verbs, the correspondence between what a model predicts and what speakers actually do is not always perfect. This experiment seeks to answer whether native speakers can, and do, attend to the strongly-predictive cue of the verb pattern in generalizing existing masdar patterns to nonce forms.

4.2 Experiment 2

4.2.1 Methodology

4.2.1.1 Participants

Participants were recruited via Amazon's Mechanical Turk. The heading for the experiment was "Answer a survey about Arabic words" (in Arabic). Participants received $4 upon completion of the experiment. Participants who completed experiment 3 were not allowed

to participated in experiment 2 as both involve masdars, but participants who completed either experiment 1A or 1B were allowed to participate as they examine a different morphological system. In total, 146 participants accepted the task on Mechanical Turk, of which only 72 completed any experimental items. 68 participants completed the entire experiment. 28 of those participants were excluded from analysis for the following reasons: non-native speaker of Arabic (n=2), demographic information not recorded in database (n=4), or achieving less than 70% accuracy on filler items (n=22). In total, 40 participants completed the experiment and met all qualifications for inclusion in analyses.

Of the 40 participants whose data was analyzed, 24 participants were male and 13 were female. Gender was not recorded for 3 participants due to a database error; all other demographic information was recorded for those participants so the participants were included in analyses**.** All participants were self-reported native speakers of Arabic. Mean proficiency in MSA was 8.65 on a scale of 1-10 (S.D.=1.54). Mean frequency of use of MSA was 7.39 on a scale of 1-10 (S.D.=2.36), with 1 being "rarely use MSA" and 10 being "use MSA frequently." For level of education, 6 participants reported having less than a college education, 27 participants an undergraduate education, 5 participants a master's degree, and 2 participants a doctorate. 38 participants reported also speaking English, with a mean proficiency of 7.96 (S.D.=2.02) on a scale of 1-10. All participants reported speaking a second language (including English), 18 reported speaking a third, and 2 reported speaking a fourth.

Information on primary spoken dialect was also elicited. Dialects were classified by major regional dialect. 12 participants reported speaking Egyptian as their primary dialect. 4 participants reported speaking a Levantine dialect (includes Jordanian & Palestinian). 10 participants reported speaking a North African dialect (includes Moroccan & Tunisian). 7

participants reported speaking a Peninsular dialect (includes Bahraini, Emirati, and Saudi). 4 participants reported speaking a Mesopotamian dialect (Iraqi), and 3 participants did not specify a dialect. All participants were analyzed in the main results, but only dialect groups including at least 7 speakers were used in the dialect analysis.

4.2.1.2 Experimental materials

4.2.1.2.1 Stimulus design

This experiment examines speaker preference for the masdar for multiple-listing verbs with two attested masdars. The target items were 36 existing verbs that have two masdar forms listed in Wehr (1976), one of which is the dominant[12] pattern for the voweling pattern of the verb ([CaCC] for [CaCaCa] verbs, [CaCaC] for [CaCiCa] verbs, and [CaCaaCaT] for [CaCuCa] verbs). The verbs varied in the non-dominant patterns, with 10 non-dominant patterns in total represented in the dataset. Thus, the verbs overall took 13 different masdar patterns. The target verbs were equally split amongst the verb patterns, with 12 target verbs having each verb pattern. Mean frequency of the masdars for the target verbs was 7.24 per million (S.D.=29.61). Due to the limited number of available verbs that fit the criteria for inclusion in the experiment, masdar frequency was not controlled, and some masdars did not occur in Aralex. Mean frequency for the masdars taking the dominant pattern was 2.82 per million (S.D.=6.21), and for masdars taking the non-dominant pattern was 11.67 per million (S.D.=41.24). There was no significant

[12] I will refer to the masdar pattern that it most frequent by type overall for each verb pattern as 'dominant', as it is the dominant pattern system-wide. Thus, 'dominant' here does not refer to the masdar form that speakers prefer for a particular verb, but rather to the overall more frequent pattern for that verb pattern.

difference in frequency between the masdars with a dominant pattern and those with a non-dominant patterns, t(36.59)=-1.27, p=0.21.

The filler items were 36 existing verbs that were matched to the target verbs for voweling pattern, phonotactic probability and neighborhood density of the root, and which took one of the masdar patterns of the matched target verb. Half of the filler items took the dominant pattern for the voweling pattern of the verb, and half took the matched non-dominant pattern, which varied across verbs. The distractor masdar for the filler items was formed on the pattern of the matched multiple-listing verb. That is, if the matched target verb took the masdar patterns [CaCC] and [CaCiiC], and the filler verb took the pattern [CaCC], the distractor was created on the pattern [CaCiiC]. Like the target items, frequency could not be controlled due to the limited number of possible items, and some masdars did not occur in Aralex. The average frequency for the filler masdars was 33.98 per million (S.D.=74.21). There was a marginal but non-significant difference in frequency between masdars with the dominant pattern and those with a non-dominant pattern, t(18.1)=-2.03, p=0.057.

4.2.1.2.2 Procedure

The procedure for experiment 2 was similar to that for experiment 1B. Participants saw an introduction screen, a consent form, questionnaire on language background and demographics, instructions, a practice section with 4 items, and finally the test section.

The test section consisted of 72 items, with 36 target items and 36 filler items. Participants who did not select the correct masdar for 70% of filler items were excluded. This threshold was lowered from the 80% threshold used for experiments 1A and 1B after testing

began, as it became clear that participants had generally lower accuracy on masdars than on the noun plurals. I will consider why this might be the case in the discussion of this chapter. 70% is nonetheless above the 95% confidence interval for random guessing for 36 items. Analyzed participants had a mean accuracy of 81.2% on filler items (S.D.=6.2).

Items were presented one at a time in two-sentence frames. The verb form always occurred in the first sentence, and was marked in blue. The verb was in the past tense form in all sentences, so that the vowel pattern of the verb was available to participants. For the target items, all but six verbs were in the third person masculine singular. The remaining six were either in the first person singular or the third person feminine singular. For the filler items, all but ten were in the third person masculine singular. The difference in person and gender was due to unnaturalness of the third person masculine singular for some verbs. The second sentence contained a blank that syntactically required a noun. The two masdar options were presented below the sentences, and participants were instructed to click on the form they preferred. Figure 4.1 shows an example stimulus. One sentence frame was constructed for each filler item and each target item. Order of sentence presentation and order of masdar options were randomized for each participant.

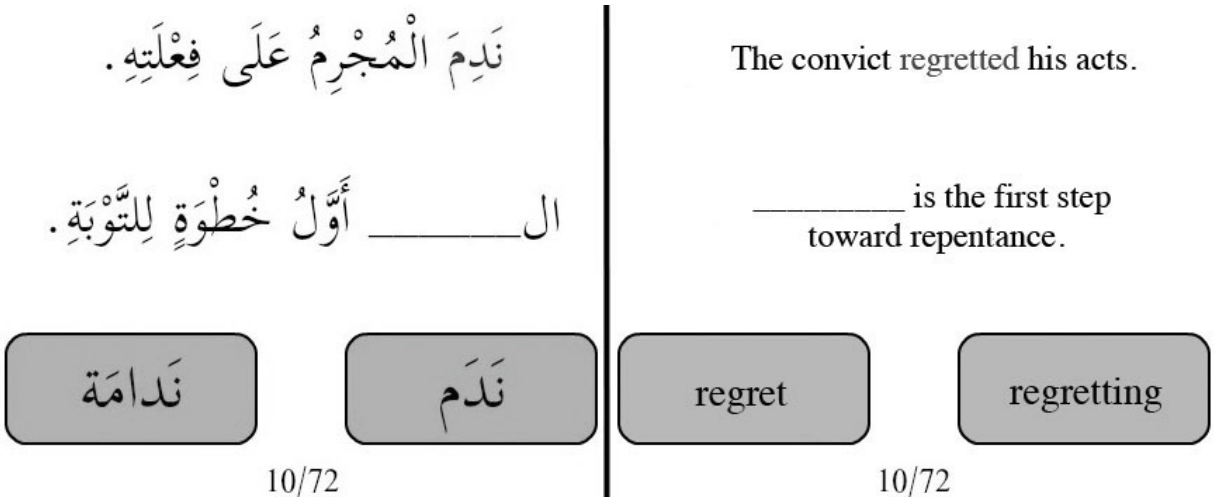

Figure 4.1: Example target item from experiment 3 (left) and English gloss (right)

4.2.2 Results

4.2.2.1 Overall results

The overall results for the multiple-listing verbs are shown in Figure 4.2. The results are arranged in order from lowest to highest proportion of responses that conformed to the default masdar pattern for the verb pattern (e.g., [CaCC] for [CaCaCa] verbs, [CaCaC] for [CaCiCa] verbs, and [CaCaaCaT] for [CaCuCa] verbs). As can be seen in the graph below, agreement overall on the preferred masdar for the multiple-listing verbs is quite variable across verbs. For some verbs, all 40 participants prefer the same form, while for other verbs, participants are almost equally split on which form they prefer. The average preference for the default pattern was 32.4% (S.D.=20.7). The average preference for the non-default pattern was 66.8% (S.D.=29.6). That is, speakers had a general preference for the non-default patterns over the default patterns, but this varied widely across verbs. If we consider agreement across speakers, which ignores whether a pattern is default or non-default, the average agreement for all verbs

was 80.7% (S.D.=13.9). Thus, for most verbs, speakers did not choose randomly between the two masdars; however, the wide differences in agreement between verbs warrants investigation.

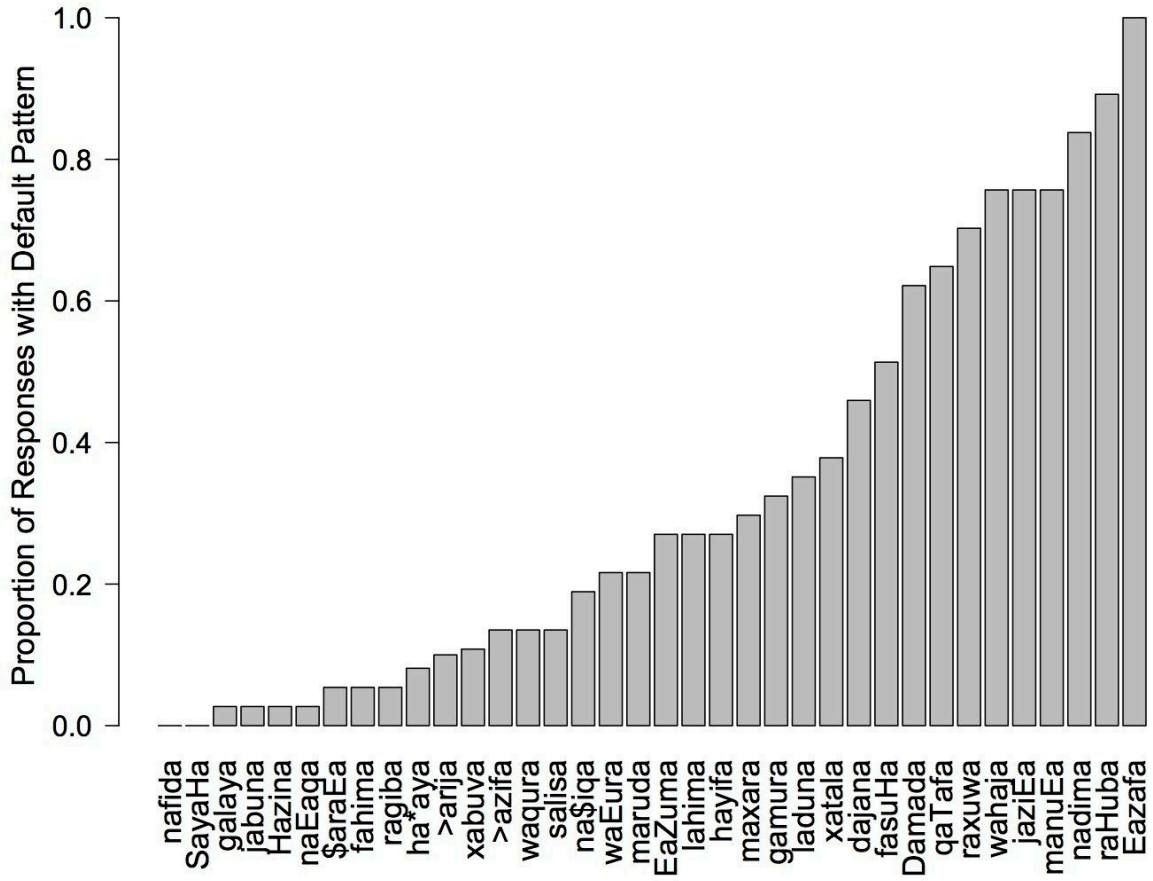

Figure 4.2: Proportion of default masdar pattern responses by item for target items

It is possible that there are differences due to the vowel pattern of the verbs, as the vowel patterns are not equally frequent by type in the lexicon. Figure 4.3 shows the proportion of default responses by item, separated into the three verb patterns. Overall, we see that speakers are less likely to select the default masdar pattern if the verb pattern is [CaCiCa], with a mean 23.6% (S.D.=27.6) of responses having the default pattern. The proportion of responses taking

the default pattern is similar for [CaCaCa] and [CaCuCa] verbs, with a mean of 36.3% (S.D.=33.8) and 37.6% (S.D.=27.9) of responses having the default pattern, respectively. Although this is interest, it is unclear why this difference would occur only for [CaCiCa] verbs. These differences may be explained by other factors, for instance frequency of the two masdars, as this was not equal across the verb patterns.

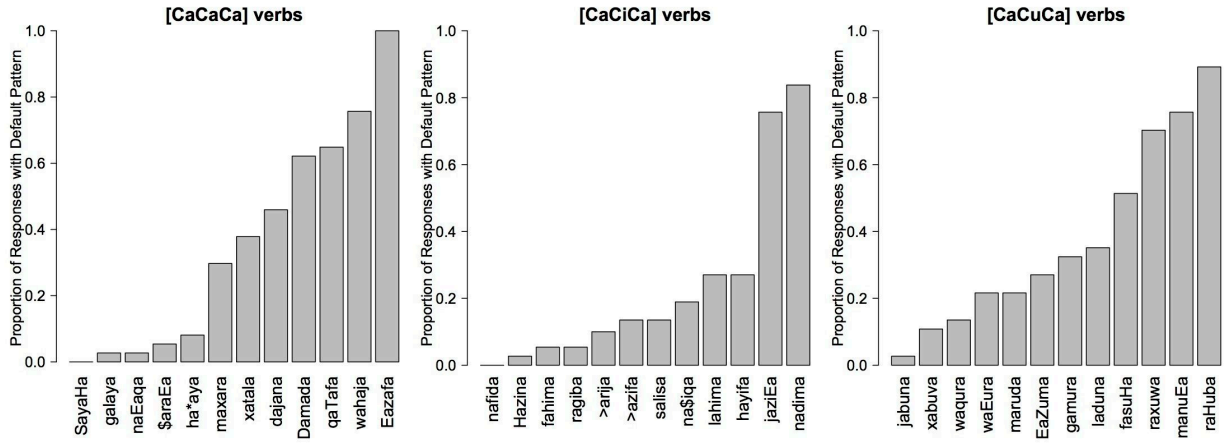

Figure 4.3: Proportion of default masdar pattern responses by item for target items, by verb pattern

One possible reason why agreement across speakers on the preferred masdar may be absolute for some forms and at chance for others is the relative token frequency of the two masdars. As noted in the methodology section, due to a limited number of items fitting the criteria for inclusion in the experiment, token frequency of the masdars was quite variable across the items. For some verbs, there was one masdar that was much more frequent than the other, while for other verbs, both masdars were equally frequent, or equally infrequent. Frequency can, roughly, be taken as an estimate of how likely a speaker is to know a given form. If one masdar for a verb is much more frequent than the other, then speakers are more likely to know it, and to thus prefer it in this type of task. However, all other factors being equal, if speakers are equally

familiar or equally unfamiliar with both masdars for a given verb, then they should select between them randomly. Thus, it is possible that agreement is largely a factor of the frequency of the two masdars.

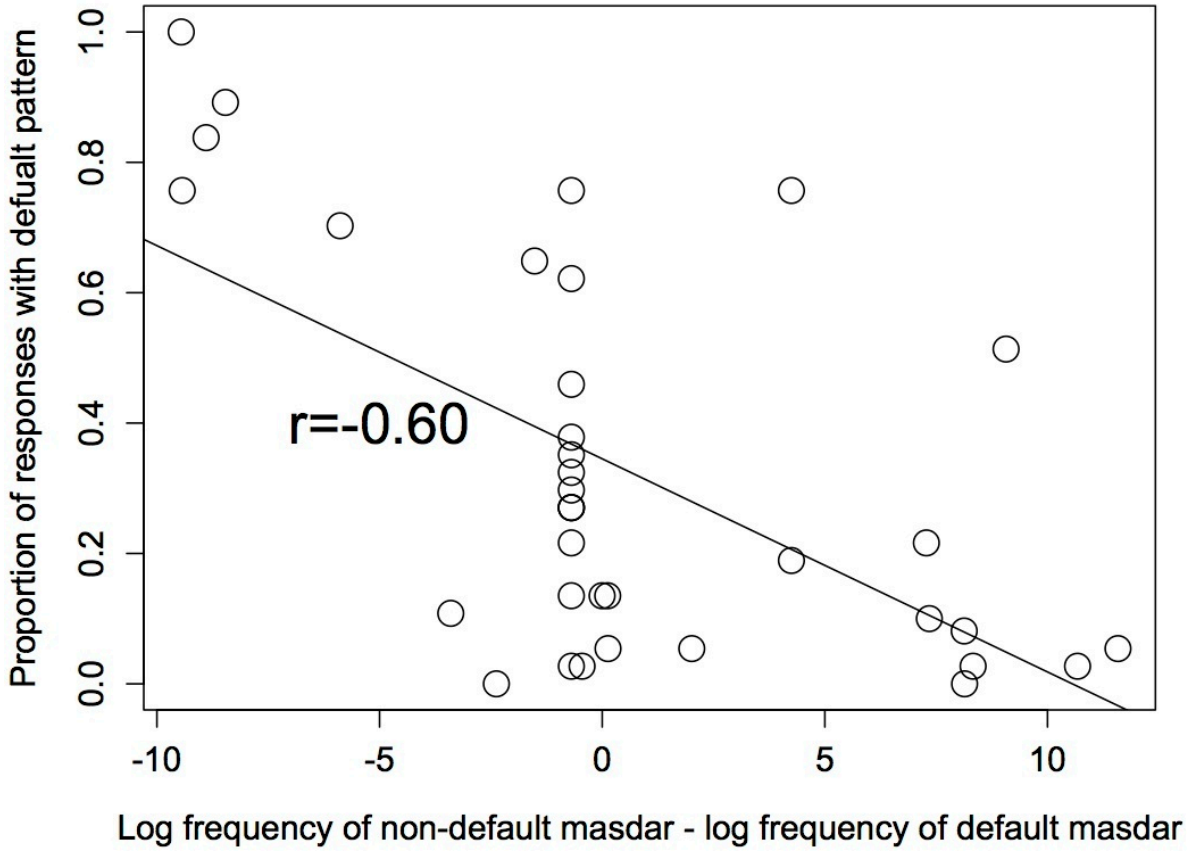

Figure 4.4: Difference in log frequency of non-default and default masdar vs. proportion of responses with default pattern

However, the effect of relative frequency of the two masdars on masdar preference is relatively weak, as shown above in Figure 4.4. There is a negative correlation between the frequency of the non-default masdar minus frequency of the default masdar, which indicates the extent to which the non-default form is dominant, and speaker preference for the dominant pattern, but the correlation is not significant, r=-0.60, p=0.06. The dominance of one of the two forms in the lexicon does account for some, but not all, of the variability in responses, so there

may be other reasons for the differences observed. Based on Figure 4.4 above, it is clear that this correlation is largely driven by a few verbs where there is a large frequency difference between the two masdars. The wide spread in agreement among verbs with similar-frequency masdars is unexplained by this analysis alone.

One interesting sub-pattern to note is that concerning verbs where the non-default pattern is the system-wide default [CaCC]. There were four verbs that fit this criteria, all having the pattern [CaCiCa] and taking the two masdar patterns [CaCaC] and [CaCC]. In all four cases, the non-default [CaCC] was preferred over what should have been the preferred pattern for that verb pattern. On average, speakers preferred the [CaCC] pattern 66.5% of the time, even though frequency of the two masdars was nearly equal for three of the verb pairs. There is no other non-dominant masdar pattern that speakers show a general preference for. Interestingly, though, when the verb pattern was [CaCaCa], speakers did not show the same preference for the default form [CaCC], with some verbs having nearly 100% preference for the non-default pattern and others having nearly 100% preference for the default pattern. It is nonetheless interesting that speakers would prefer the system-wide default for the four verbs noted above even when the competing pattern is statistically dominant, and the frequencies of the masdars are roughly equal.

The best way to assess whether these effects are significant when considered in aggregate is using a linear mixed effects model. In this way, the interesting patterns noted above can be confirmed (or disconfirmed) statistically. A model was constructed which tried to predict whether a participant would select the default or non-default pattern for a particular item using the following fixed effects: log token frequency of the default masdar, log token frequency of the non-default masdar, probability of the default masdar pattern given the verb pattern, and

probability of the non-default masdar pattern given the verb pattern. In addition, a random effect of participants was included. No interactions between factors were included.

A second model was run which was identical to the first, except that the probabilities of the masdars were the probabilities of the masdar overall in the system rather than by the verb pattern. By comparing these two models and seeing which better predicts responses, we can ascertain whether participants are using type statistics on the verb pattern in forming judgments on the test items, or type statistics on the entire masdar system. The modeling results in experiment 2 show that the optimal strategy for predictiveness is to use type statistics on the verb pattern; however, comparing these two models allows us to test whether participants line up with the best-predictive models, as we have previously seen that this is not always the case.

First, we find that the model using pattern probabilities on the overall system has better explanatory power than the model using pattern probabilities on the verb pattern. This was assessed using the AIC, BIC and log likelihood from the model summaries. Next, we will examine the individual factors in the better-fitting model, which used overall pattern type statistics rather than types statistics on the verb pattern, to assess which fixed effects are significant predictors of whether a participant will select the default masdar pattern. Significance for each factor was determined used nested model comparison (Barr et al., 2013). First, the log frequency of the default masdar was a significant positive predictor of selecting the default pattern, β=0.12, S.E.=0.02, $\chi^2(1)$=55.27, p<0.001, and the log frequency of the non-default masdar was a significant negative predictor of selecting the default pattern, β=-0.21, S.E.=0.02, $\chi^2(1)$=187.24, p<0.001. The probability of the default masdar pattern was not a significant

predictor of selecting the default masdar, β=0.07, S.E.=0.24, $\chi^2(1)$=0.09, p=0.76, nor was the probability of the non-default masdar pattern, β=-0.69, S.E.=0.37, $\chi^2(1)$=3.68, p=0.055.

4.2.2.2 Filler results

As noted above, participant accuracy on filler items was overall lower in this experiment than in the noun plural experiments. In addition, accuracy across filler items was highly variable. There were some verbs for which all 40 participants selected the correct masdar, but there were also some verbs for which fewer than half of participants selected the correct masdar; for example, for one item, only 42.5% of participants selected the correct masdar. Figure 4.5 shows accuracy by item for all filler items, in ascending order. Mean accuracy for the filler items was 80.7% (S.D.=19.4), with 30 of the 36 items having greater than 60% accuracy across participants. The variability in accuracy is quite interesting, given that these are all existing verbs with a single attested masdar.

One possibility is that participants are more likely to be accurate on items that conform to the default patterns for the verb pattern, as they have more type support for those items. However, there is no significant difference in accuracy between verbs taking a default masdar pattern and those taking a non-default masdar pattern, t(33.87)=-0.99 , p=0.33.

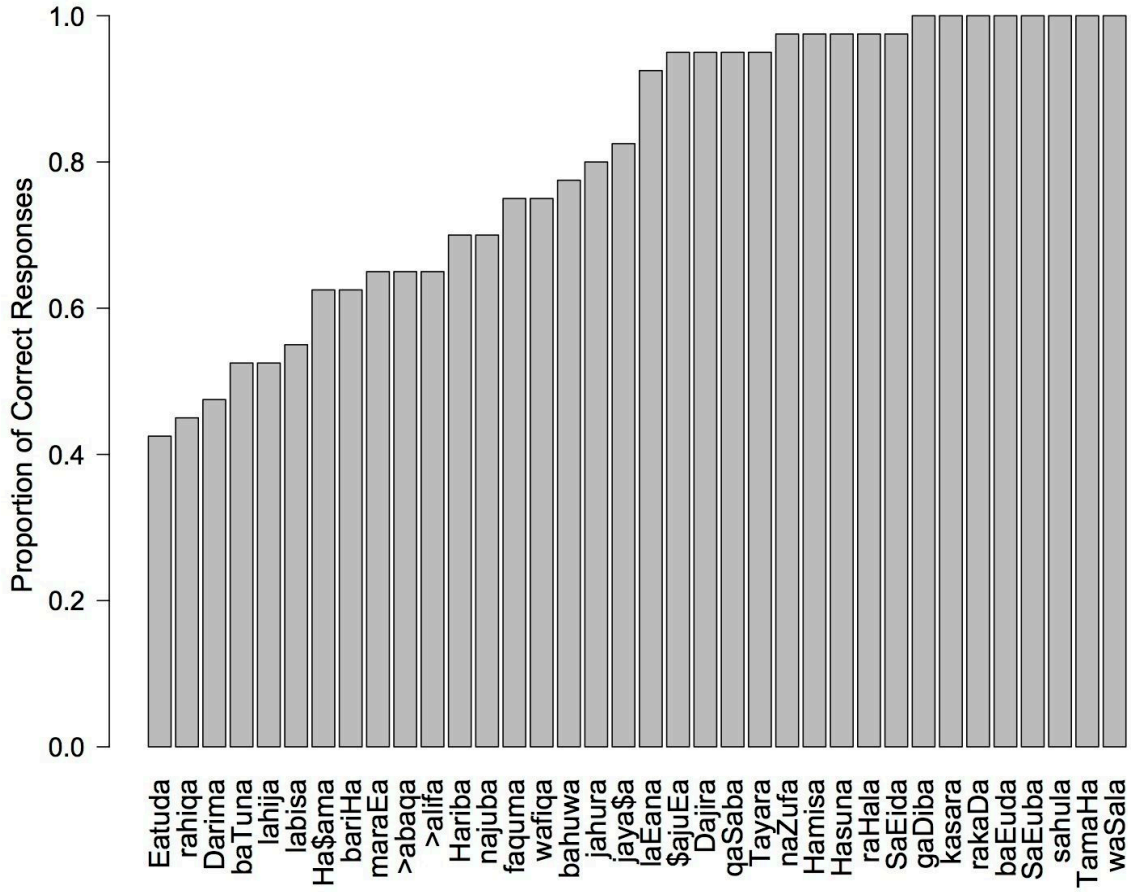

Figure 4.5: Proportion of correct responses by item for filler items

Another possibility is that there may be differences in accuracy across the verb patterns. Figure 4.6 shows the proportion of correct responses by item, separated by verb pattern. For [CaCaCa] verbs, the mean accuracy is 87.9% (S.D.=15.1), while for [CaCiCa] verbs it is 71.9% (S.D.=20.8) and for [CaCuCa] verbs it is 82.3% (S.D.=19.7). Thus, we see a similar pattern to above, with participants showing lower accuracy on [CaCiCa] verbs, and similar accuracy on the other two verb patterns. However, as noted, token frequency of the masdar is not equal across the verb patterns, and this may be a confounding factor.

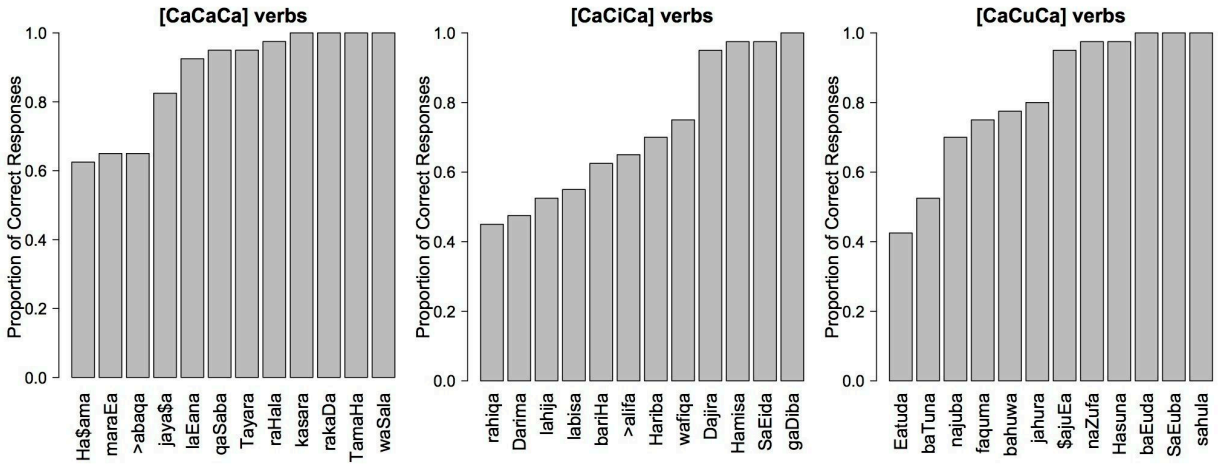

Figure 4.6: Proportion of correct responses by item for filler items, by verb pattern

As mentioned above, participants may be more familiar with, and thus more accurate, on more frequent masdars, which is a separate issue from the default or non-default pattern, or the verb pattern. There is a significant positive correlation between the log frequency of the masdar and accuracy, $r=0.40$, $p<0.05$. Thus, for filler items, speakers are more accurate at identifying the masdar if it is more frequent. Overall, though, participants show relatively lower accuracy on filler masdars for this experiment than on filler noun plurals in experiment 1A. The possible reasons for this will be discussed in the discussion.

In order to ascertain what other factors might be affecting participant accuracy on the filler items, linear mixed effects models similar to those used on the test items were constructed. The models reported here tried to predict whether a participant would choose the correct masdar for a given item using the following fixed effects: token frequency of the correct masdar, token frequency of the distractor masdar (as some did exist as real words, but not as masdars for that verb), probability of the correct masdar pattern given the verb pattern, probability of the

distractor masdar pattern given the verb pattern, and whether the correct masdar had a default pattern. In addition, a random effect of participants was included. No interactions were included.

A second model was constructed which was identical to the first, but using masdar pattern probabilities for the overall system rather than by the verb pattern. Like with the test items, we find that the model using masdar pattern probabilities on the overall system has slightly better explanatory power than the model using masdar pattern probabilities on the verb pattern, again assessed using the AIC, BIC and log likelihood from the model summary. As above, we will examine the individual factors in the better-fitting model, which used overall pattern type statistics rather than types statistics on the verb pattern, to assess which fixed effects are significant predictors of whether a participant will select the correct masdar used nested model comparison (Barr et al., 2013).

First, the log frequency of the correct masdar was a significant positive predictor of selecting the correct masdar, β=0.10, S.E.=0.015, $\chi^2(1)$=50.00, p<0.001. The log frequency of the distractor item was also a significant positive predictor of selecting the correct masdar, β=0.05, S.E.=0.026, $\chi^2(1)$=4.82, p=<0.05. The probability of the correct masdar pattern was not a significant predictor of selecting the correct masdar, β=0.37, S.E.=0.31, $\chi^2(1)$=1.45, p=0.22, nor was the probability of the distractor masdar pattern, β=-0.67, S.E.=0.34, $\chi^2(1)$=3.82, p=0.05. Finally, whether the correct masdar had a default pattern was a significant positive predictor of selecting the correct masdar, β=4.02 S.E.=0.15, $\chi^2(1)$=7.03, p<0.01.

4.2.2.3 Analysis of dialect background

In the overall results, there was very low agreement between speakers on the masdar for some verbs. As noted in chapter 3, one possibility for the existence of verbs with multiple masdars is that the masdar for a given verb is dialectally variant. The low agreement observed in this experiment for some verbs could stem from different dialects using different masdars for a given verb. This section will examine whether the dialect background of the participants in this experiment has a reliable effect on the masdar forms they prefer.

Participants were classified by major dialect region. The dialect regions examined here are: Egyptian (n=12), North African (n=10) and Peninsular (n=7). The proportion of speakers choosing the default masdar pattern for each target item is shown in Figure 4.7, with items arranged by proportion of speakers across all dialects selecting the default pattern. There are some minor differences between dialects, but there are no particular items or regions that stand out as being hugely variant across dialects. However, this needs to be confirmed statistically. Krippendorff's alpha (Krippendorff, 1980) was used to measure inter-speaker agreement within each dialect group and across all dialect groups. This coefficient computes the overall agreement above chance between raters on assigning *n* items to *c* categories, where 0 = no agreement at all and 1 = perfect agreement. In the case of this experiment, a 'rater' is a participant, an 'item' is a multiple-listing verb, and a 'category' is a masdar pattern.

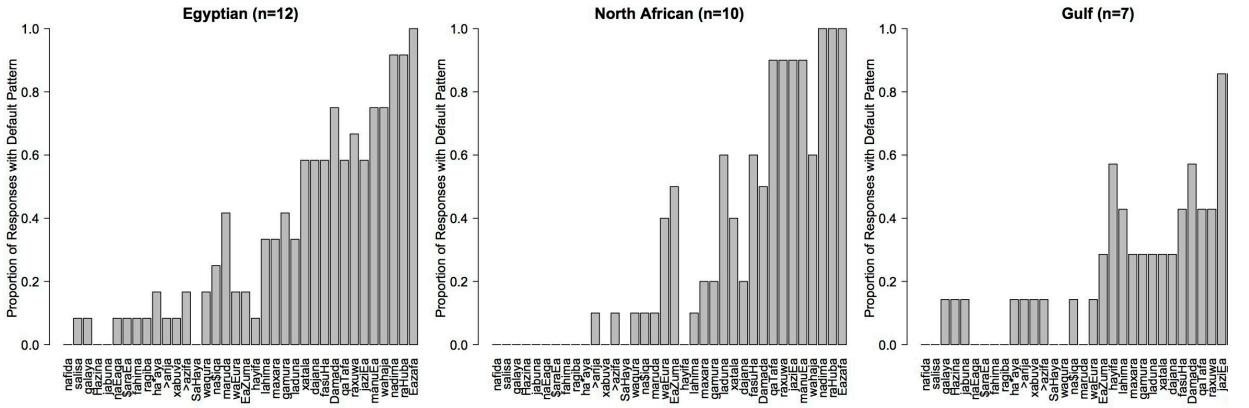

Figure 4.7: Proportion of default masdar pattern responses by item, by dialect group

Across all dialects, α=0.684. Within each dialect group, the agreement is similar to that across all participants: Egyptian, α=0.665; North African, α=0.785; Peninsular, α=0.689. The overall level of agreement is slightly higher for the North African group than across all speakers, but agreement for the other dialect groups is similar to overall agreement. A similar pattern was observed in experiment 1A, wherein North African speakers showed slightly higher inter-speaker agreement than the other dialect groups. However, in both cases, the differences are very minimal, and thus no firm conclusions can be drawn. For the filler items, this pattern of agreement is very similar. Across all dialects, α=0.715. Within each dialect group, agreement is similar to across dialect groups: Egyptian, α=0.725; North African, α=0.74; Peninsular, α=0.698. Thus, the pattern of accuracy on particular filler items does not seem to be dialect-related. Overall, it does not appear to be the case that there is overall higher agreement between speakers from the same dialect background than between speakers from different dialect backgrounds on either test or filler items. This suggests that the hypothesis of multiple masdars for a single verb stemming from different dialects having different masdars is false. In order to confirm this, larger numbers of participants would be necessary. Nonetheless, the minimal differences

between speakers of different dialects in this experiment strongly suggest that the (dis)agreement patterns across speakers stems from some other source than dialect background.

4.2.3 Discussion

In summary, I have demonstrated that many of the verbs in this experiment do not truly have two active masdars, with participants showing about 80% agreement across verbs on the preferred masdar of the two options. Second, I have shown that for verbs with low agreement across speakers, the pattern of agreement cannot be explained by the dialect background of the speakers. Thus, for verbs that may have two active masdars (meaning those that have low agreement across speakers), this does not stem from lexical differences across dialects. Finally, I have shown that Arabic speakers do not know the masdars for many verbs as well as would be predicted. In comparison to noun plurals of similar frequency, Arabic speakers achieve much lower accuracy in identifying masdars for the filler verbs.

For individual predictors of selecting the default masdar pattern, we find an interesting mix of effects. For the test items, there was a general overall tendency to select the non-default pattern. In addition, the log frequencies of both masdars were significant predictors. The log frequency of the default masdar was a positive predictor of selecting the default pattern, while the log frequency of the non-default masdar was a negative predictor of selecting the default masdar. However, neither the probability of the default masdar pattern nor the probability of the non-default masdar was a significant predictor, which is surprising.

For the filler items, we find a similar pattern. Participants showed a general tendency toward accuracy, although as noted they generally had lower accuracy than participants in the

noun plural experiments. The log frequency of the masdar was a significant positive predictor of accuracy. In addition, the log frequency of the distractor masdar (for which only a few items had non-zero values) was also a significant positive predictor of accuracy, which is unexpected. The probabilities of the correct masdar pattern and distractor masdars pattern were not significant predictors. Finally, whether the masdar had a default pattern had a positive effect on accuracy.

For both the test and filler items, the effect of token frequency of the masdars is not very surprising. For the test items, the relative frequencies of the two masdars should indicate the extent to which a speaker is familiar with the two masdars, and speakers should be biased toward the one they are more familiar with. Likewise, for the filler items, token frequency is a good indicator of whether they know the masdar for that verb. The positive effect of the distractor masdar frequency is somewhat surprising. One possibility for this effect is that it indicates familiarity with the morphological paradigm of the filler verb. It is also possible that participants were more likely to know that higher-frequency distractors were not masdars, and thus were more easily able to rule out the distractor as the masdar. Either of these possibilities would lead to higher accuracy on the filler item.

The lack of effects of pattern probability was somewhat surprising. It is unclear why this would be the case, but it may be a task effect, since participants saw a wide variety of relatively rare masdar patterns, and thus the individual token frequencies of the items may have been the driving force behind preference for test items and accuracy for the filler items. One significant effect observed for the filler items, however, was that participants were more likely to be correct on masdars with default patterns. To some extent, this suggests that pattern type frequency plays a role, as default patterns have significantly more type support than non-default patterns.

However, for the test items, participants showed a general tendency to select the non-default pattern, which is inconsistent with this suggestion. Given the relatively small number of test and filler items in this task (n=36 for each), and the variation in both frequency and pattern probability across items, we may not be able to draw many conclusions about these apparent inconsistencies.

With regards to the test verbs with low agreement across speakers, there are two possible sources of this variation, which may be at least in part explained by the factors examined above. The first is that some of the masdars in the experiment are currently undergoing leveling. If this is the case, then speakers should be familiar with both masdars, but may not have a strong aggregate preference for one form over the other. The level of preference would depend on how far the leveling has progressed, which may be generally estimated by token frequency of the two masdars, and we do find significant effects of the frequencies of the two masdars on which masdar participants select. Interestingly, however, the directionality of the leveling is not what one would expect, with participants showing a general preference for the non-default masdar patterns. There are certainly documented cases of analogical leveling wherein the pattern has been leveled to a non-default, as in [dive] ⇒ [dived] versus the newer [dove], but these are rarer (at least in English) than leveling to the default or dominant pattern, for instance as in [weave] ⇒ [weft] versus the newer [weaved] (e.g., Deutscher, 2001). The general tendency toward selecting the non-default patterns, in fact, suggests that there may be analogical leveling to the non-dominant patterns for some of the items in this experiment.

The second active possibility, which is not mutually exclusive from the first, is that speaker judgments on the multiple-listing verbs may not reflect analogical leveling, but rather general unfamiliarity with either of the masdars. In fact, overall participant agreement on the

masdar for the multiple-listing verbs is not significantly different from participant accuracy for the single-listing filler items, t(63.59)=0.017, p=0.98. This suggests that participants may be using the same strategy for both the multiple-listing verbs and the filler verbs, and in fact, be simply somewhat uncertain about the identity of the masdar for many verbs, regardless of whether the dictionary claims it has a single masdar or multiple ones. There are some differences in the significance of particular predictors between the multiple-listing verbs and the single-listing verbs that suggest that this may not be the case, however. For the test items, the frequencies of both masdars were significant predictors of selecting the default pattern, with the frequency of the default masdar a positive predictor and the frequency of the non-default masdar a negative predictor. For the filler items, however, the frequency of the distractor masdar was a significant positive predictor of selecting the correct masdar, which is the opposite direction of what would be expected. This could indicate that speakers were familiar with the distractor item and knew that it was not a masdar, which would then drive them toward the correct choice. This to some extent indicates a difference in treatment between the test and filler items. The second major difference is in the tendency to select the non-default patterns for the test items versus the tendency to select the default patterns for the filler items. It is not entirely clear why this difference would occur, in particular given that the filler and test items were balanced such that half took a default pattern and half a non-default pattern. Overall, these differences in predictors between the test and filler items suggest that this second hypothesis is not correct. Then, the more likely possibility is that speakers are aware that there are two live candidate masdars for the test items, and that the results for the test items are most consistent with the leveling hypothesis.

In summary, we find that for many verbs that are claimed to have two active masdar forms, speakers show a general preference for one. This is an important take-home point, as it

suggests that the prevalence of verbs with multiple masdars is overstated in the literature and in dictionary sources (e.g., Wehr, 1976; Wright, 1988). As I have discussed previously, dictionary sources are often outdated, and this result suggests a strong need to keep Arabic dictionaries better updated, in particular with regards to the masdars of form I verbs. Further, the finding that the token frequencies of the two masdars, which should generality indicate speaker familiarity with the masdars, are significant predictors of which masdar speakers prefer, is consistent with the hypothesis that many of these verbs are currently undergoing or have undergone leveling.

One major unanswered question is if, and how, speakers extend masdars to unseen forms, which is an important aspect of studying learnability. If the system is truly learnable on the basis of the type statistics on the verb pattern, then we would expect participants in a nonce-form task to generalize masdars in a manner that reflects this. In the next experiment, I will examine how speakers generalize existing masdar patterns to nonce verbs, and on the basis of this task, we can draw more firm conclusions about how well speakers truly know the masdar system and the subregularities within. With these results, we can better disentangle the possible differences in accuracy on these tasks, and in learnability in general, between the noun plural and masdar systems.

4.3 Experiment 3

4.3.1 Introduction

As noted in the previous discussion, based on the results of experiment 2, speakers of Arabic seem to have a weaker knowledge of the masdar system than would be expected based on

the modeling work in chapter 3. The masdar for an unseen verb is be about 83% predictable using type statistics on the verb pattern. This is an important comparison to the noun plural system, where is it clear that speakers are able to learn type statistics on the coarse-grained phonological generalization of the CV template, and extend noun plurals to nonce forms in a manner that reflects these statistics. Thus, we would expect that speakers of Arabic should also be able to learn type statistics on the coarse-grained phonological representation of the verb pattern. If this is the case, then speakers in a nonce-form experiment should also generalize masdar patterns in a manner that reflects the type statistics on the verb pattern, whether by primarily generalizing the dominant pattern for that verb pattern, or by matching type statistics on the verb pattern in generalization.

A second open question is whether speakers will generalize masdar patterns in a probabilistic or a deterministic manner. In experiments 1A and 1B, participants generally employed a probability-matching strategy, but the noun plural system differs from the masdar system in several key ways. First, the predictability of each system as a whole is different, with the masdar overall more predictable than the noun plural system in analogical modeling. This may lead speakers to select among available patterns more deterministically, as they should have higher certainty about what the masdar for an unseen form is. However, as seen in experiment 2, native speakers of Arabic actually have lower certainty about the masdar for individual verbs than about the plural for singular nouns of comparable token frequency. This is a different type of uncertainty that overall system predictability, but one that may be quite relevant for the experiment at hand. In experiment 3, I will examine these two questions using a nonce-form task in which speakers are asked to create a masdar for an unseen but Arabic-like verb in an open-response paradigm. By examining generalization of nonce verbs that speakers have never

encountered, we can avoid some of the confounding factors of token frequency of individual items and examine how well speakers know and can generalize the patterns in the system as a whole. In addition, the open-response paradigm places no limitations on what masdar patterns speakers are able to use, which allows us to examine how speakers navigate generalization when many possibilities are active.

4.3.2 Methodology

4.3.2.1 Participants

Participants were recruited via Amazon's Mechanical Turk. The heading for the experiment was "Answer a survey about Arabic words" (in Arabic). Participants received $5 upon completion of the experiment. Participants who participated in experiment 2 were blocked from taking part in experiment 3. In total, 441 participants accepted the task on Mechanical Turk, of which only 67 completed any experimental items. 51 participants completed the entire experiment. 10 of those participants were excluded from analysis for the following reasons: non-native speaker of Arabic (n=3), previously completed experiment 2 (n=1), or not having fully diacritized responses or responses with licit syllable structure for at least 80% of forms (n=6). In total, 41 participants completed the experiment and met all qualifications for inclusion in analyses.

Of the 41 participants whose data was analyzed, 22 participants were male and 16 were female. Gender was not recorded for 3 participants due to database error; all other demographic information was recorded for these participants so they were included in analyses**.** All participants were self-reported native speakers of Arabic. Mean proficiency in MSA was 8.63 on

a scale of 1-10 (S.D.=1.73). Mean frequency of use of MSA was 6.98 on a scale of 1-10 (S.D.=2.58), with 1 being "rarely use MSA" and 10 being "use MSA frequently." For level of education, 4 participants reported having less than a college education, 23 participants an undergraduate education, 9 participants a master's degree, and 5 participant a doctorate. 39 participants reported also speaking English, with a mean proficiency of 8.74 (S.D.=1.39) on a scale of 1-10. 17 participants reported speaking a third language, and 6 reported speaking a fourth.

Information on primary spoken dialect was also elicited. Dialects were classified by major regional dialect. 8 participants reported speaking Egyptian as their primary dialect. 16 participants reported speaking a Levantine dialect (includes Jordanian, Palestinian, Lebanese, & Syrian). 9 participants reported speaking a North African dialect (includes Moroccan, Libyan, & Tunisian). 3 participants reported speaking a Peninsular dialect (includes Saudi). 2 participants reported speaking a Mesopotamian dialect (Iraqi), and 3 participants did not specify a dialect. All participants were analyzed in the main results, but only dialect groups including at least 8 speakers were used in the dialect analysis.

4.3.2.2 Experimental materials

4.3.2.2.1 Stimulus design

This experiment uses phonotactically licit nonce verbs to examine speaker generalization of masdar patterns for the three past tense verb patterns. The target items were 180 nonce verbs that were constructed using the same procedure as in experiments 1A and 1B. Nonce roots were matched to filler roots for phonotactic probability and neighborhood density, and there were no significant differences between the nonce and filler roots for these measures. The selected roots

were cross-checked against the Aralex database (Boudelaa & Marslen-Wilson, 2010) and by a native Arabic speaker to ensure that they were not existing roots. Finally, the nonce roots were inserted in the pattern of the corresponding filler item. For instance, for the filler item [kataba], the pattern was [$C_1aC_2aC_3a$]. One matched nonce root was [**ʃmk**], resulting in the nonce singular [**ʃamaka**]. Additionally, nonce roots used in experiments 1A and 1B were excluded from experiment 3. Each participants saw one of five sets of the nonce verbs, for a total of 36 nonce items total per participant. Although verbs in Arabic are relatively closed-class compared to nouns, Frisch & Zawaydeh (2001) found that native speakers were willing to accept nonce verbs with appropriate phonotactics, and thus there is support for this methodological approach.

Participants also saw 36 filler items to ensure attention to the task. The filler items were existing verbs that take the dominant masdar pattern for that voweling pattern. Some of the filler items from experiment 2 which fit the criteria for this study were used in experiment 3. Only filler items that had a minimum accuracy of 75% in experiment 2 were used in experiment 3, and filler items re-used in experiment 3 had a mean accuracy of 91.7% in experiment 2. Because some items were re-used, participants that participated in experiment 2 were not allowed to participate in experiment 3. The filler items were also used as qualifying questions to ensure that participants were proficient in Arabic and completed the task as instructed. Filler items had a mean frequency of 12.74 per million in Aralex (S.D.=19.65). As in experiment 2, filler items had a wider range of frequency than filler items in the noun plural experiments due to the small number of verbs that fit the criteria for inclusion in the experiment. Of the 36 filler items, 11 did not occur in Aralex. Frequency will be examined as a factor in analyses of filler accuracy.

4.3.2.2.2 Procedure

The procedure for experiment 3 was similar to the procedure for experiment 1A. Participants saw an introduction screen, a consent form, questionnaire on language background and demographics, instructions, a practice section with 4 items, and finally the test section. The test section consisted of 72 items, with 36 filler items and 36 nonce items. The threshold for accuracy was different for this experiment than for the previous experiments, as experiment 2A showed overall lower accuracy for existing masdars than for existing noun plurals. In addition, participants generally have lower accuracy in open response tasks than in forced choice tasks, as the baseline for an open response task is 0%, while it is 50% for a two-way forced choice task. Thus, there was not a specific numeric threshold of accuracy for this experiment. Rather, in order for the data for a given participants to be included in analyses, at least 80% of filler and nonce items had to have licit Arabic phonotactic structure that was transparent in the orthography. Thus, if a participant did not fully vowel 80% or more of their responses, or if fewer than 80% of responses conformed to Arabic syllable structure, the participant was excluded from analysis. Because accuracy varied across participants, this factor will be analyzed separately in the results. Overall, analyzed participants had a mean accuracy of 59.7% on filler items (S.D.=14.3).

Items were presented one at a time in two-sentence frames. The verb always occurred in the first sentence in the imperfective form such that the verb pattern was transparent, and was marked in blue (displayed as grey below). The second sentence contained a blank that syntactically required a noun. Below the sentences, there was a text input box where participants typed in what they would put in the blank in the second sentence. Participants could not continue to the next page until they entered text with at least one short vowel diacritic. An example stimulus is shown in Figure 4.8.

Two sentence frames were created for each filler item. Filler items were always presented in one of the two matched sentence frames, to ensure that the verb appeared with a semantically appropriate frame. The sentences frames were balanced across participants. Nonce items were randomized for sentence frame in order to control for possible effects of semantics introduced by the sentence frame. Order of presentation was randomized for each participant.

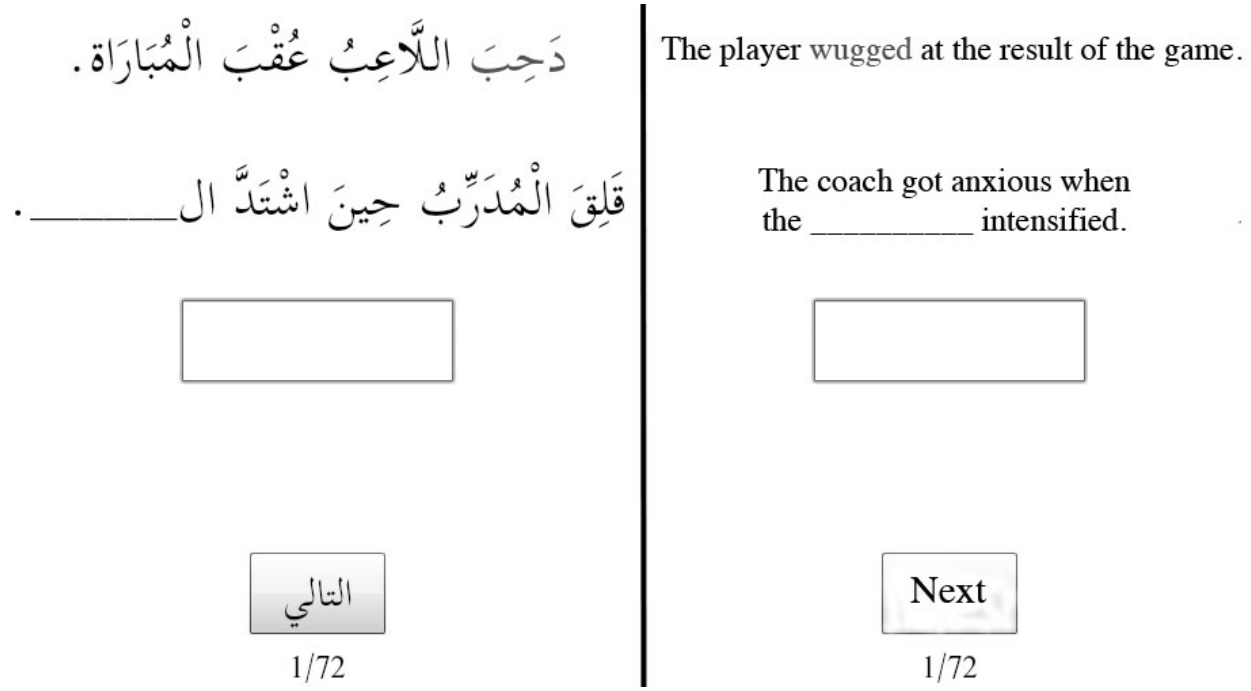

Figure 4.8: Example nonce item from experiment 3 (left) and English gloss (right)

4.3.2.2.3 Response coding

Participant responses to test items were coded for masdar pattern, CV template, whether the masdar pattern occurred in the set of existing masdars for the verb pattern (expected vs. unexpected), and the ranking of the masdar pattern in the set of existing forms. Filler items were coded for accuracy, as well as for masdar pattern.

The procedure used for stemming and pattern conversion was similar to that in experiment 1A. Responses were stemmed for: the definite article [al], possessive pronoun suffixes, and case markings. The responses were then converted to pattern, such that consonants occurring in the singular were replaced by C but vowels and additional consonants in the masdar pattern were maintained (ex: [ʃaxafaan] ⇒ [CaCaCaan] from verb [ʃaxafa]). These were checked by hand to ensure that the patterns had licit syllable structure.

With regards to diacritization, a similar procedure to that used in experiment 1A was used in experiment 3. If a base consonant did not have a diacritic, it was assumed that it was unvoweled if it was a consonant. If it was [w], [j] or [ʔ]/[A], it was assumed to be a long vowel if the character did not occur in the nonce verbal root. If it did occur in the nonce verbal root, and it was in the correct order relative to the other verbal root consonants, it was assumed to be a consonant. Additionally, [h] in final position was considered to be [T] if and only if there was a short [a] on the previous character and [h] was not in final position in the verbal root.

The pattern for each response was converted to CV template by replacing short vowels with [v] and long vowels with [vv]. Although the pattern seems to be the primary phonological representation for the masdar system based on the modeling work in the previous chapter, the CV template will also be examined to confirm whether it is relevant to masdar formation.

Individual responses were excluded from analysis for the following reasons: metathesis of verbal root consonants (ex: [faʕaza]⇒[fuzuuʕ]), substitution of verbal root consonants (excluding those additional pattern consonants in attested masdar patterns), and illicit syllable structure. In total, 3.1% of responses were excluded from analysis (n=44).

4.3.3 Results

4.3.3.1 Overall results

The overall results for all nonce verbs are shown in Figures 4.9 and 4.10. One thing that is immediately obvious from the figure is that the responses show a very heavy-tailed distribution, and are dominated by [CaCC], with 59.4% of responses having this pattern. The second most frequent pattern is [CuCuuC], interestingly, which is not one of the three dominant patterns as defined by the verb patterns. This pattern accounts for 9.6% of responses. The third most frequent pattern is [CaCaC] (8.0% of responses), fourth-most frequent is [CaCCaT] (4.8%), and fifth-most frequent is [CuCC] (3.9%). It is not until the sixth-most frequent pattern that we encounter [CaCaaCaT] (3.1%), which is the dominant pattern for [CaCuCa] verbs. There are a total of 41 patterns in the responses, with the vast majority of them (n=31) having fewer than 10 responses per pattern, out of a total 1403 responses.

If we compare this figure to the distribution of masdar patterns in the dictionary dataset from chapter 3, shown in Figures 4.11 and 4.12, there are many striking commonalities. The most frequent pattern by far in both cases is [CaCC], and [CuCuuC] and [CaCaC] are in second or third position in both the corpus and the experimental data. In addition, [CaCaaCaT] is in fifth position in the experimental data, and fourth position in the corpus data. One notable different is the relatively frequent use of [CaCCaT] in the experiment, which is 10th-most frequent in the corpus dataset. In addition, the experimental responses include a wider range of forms than is found in the corpus set. Some of these are phonotactically licit, though unattested patterns, while others are attested patterns for other verb forms. For instance, [taCCiiC] is the dominant masdar pattern for form II verbs, and [>inCiCaaC] is the dominant masdar pattern for form VII verbs. It is not clear why participants used masdar patterns from other verb forms, but they are very

infrequent in the experimental data. The primary reason for the larger number of masdar patterns in the experimental data seems to be vowel changes to existing masdar patterns. For example, we see responses in the experimental data such as [CaCaCaT], which differs from the existing pattern [CaCaaCaT] only by vowel length in the second syllable, as well as responses such as [CuCiC], which differs from the existing pattern [CuCuC] by the vowel quality in the second syllable.

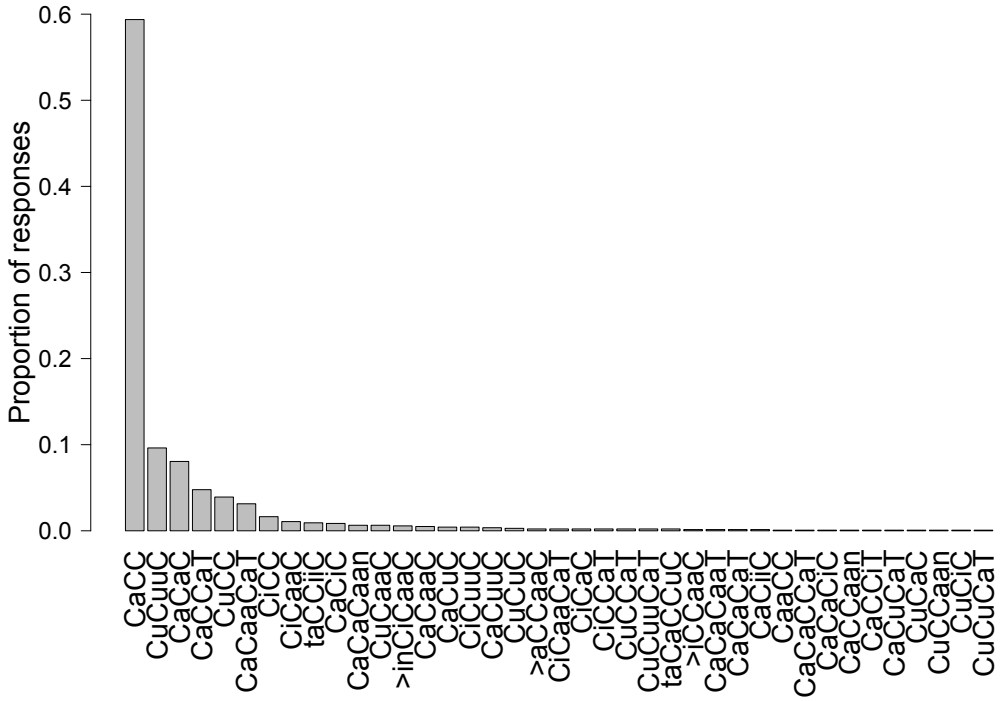

Figure 4.9: Overall masdar pattern responses, in decreasing order

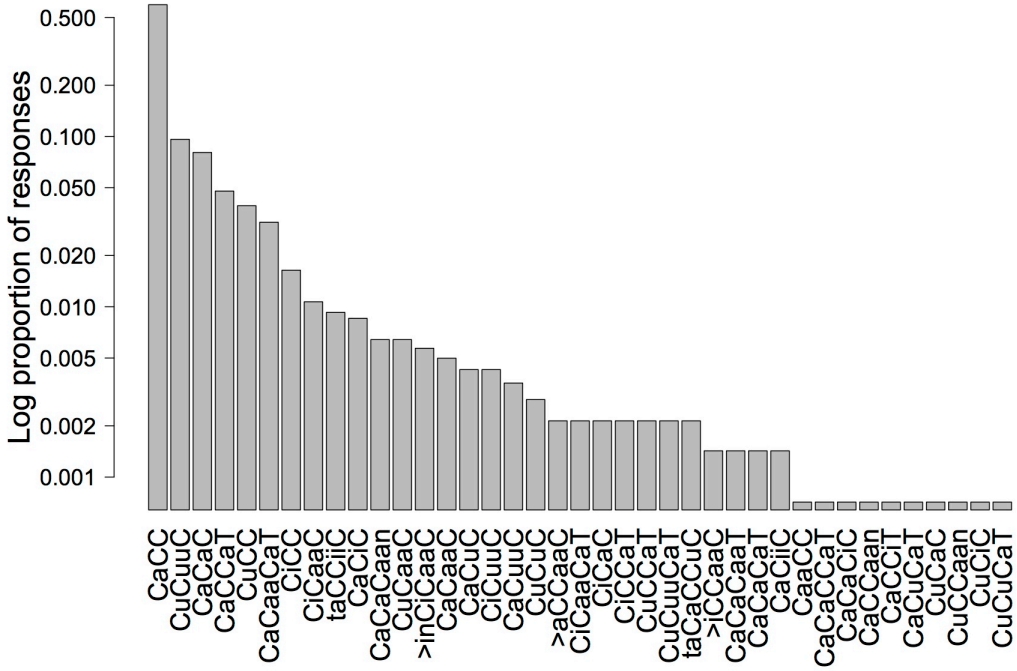

Figure 4.10: Overall masdar pattern responses, in decreasing order (log scale)

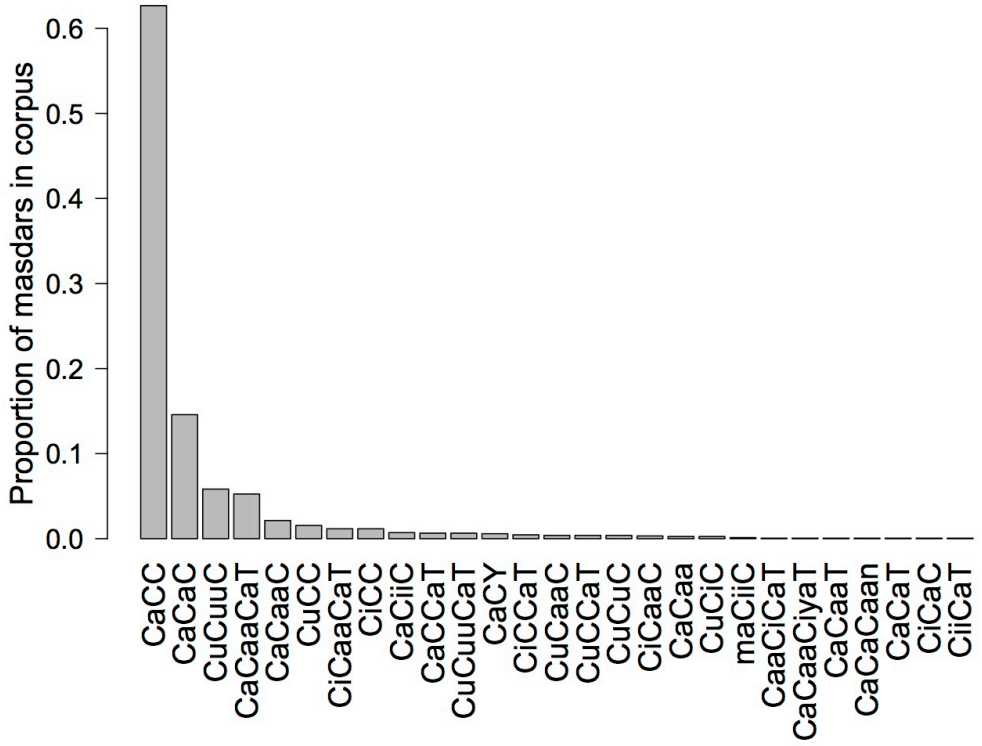

Figure 4.11: Overall corpus masdar patterns, in decreasing order

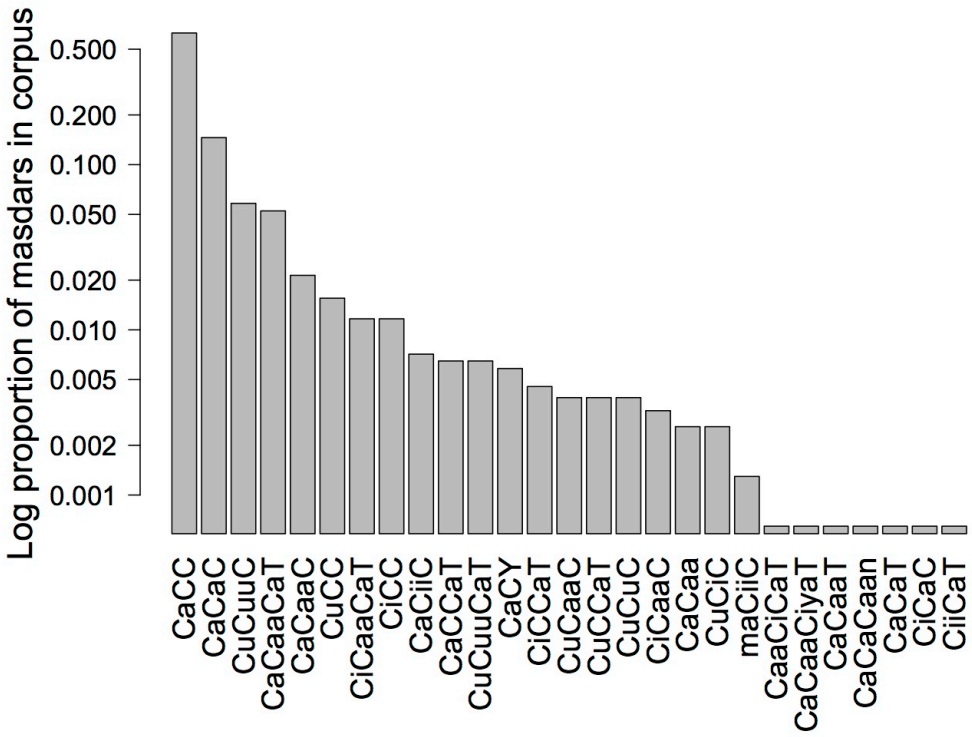

Figure 4.12: Overall corpus masdar patterns, in decreasing order (log scale)

If we directly compare the log probability of a masdar pattern in the overall corpus dataset (expected probability) to the log probability in the experimental responses, we see a significant positive correlation of $r=0.461$, $p<0.001$, as shown in Figure 4.13. The expected and observed probabilities were corrected using the Laplace correction (Lidstone, 1920) for any masdar patterns that did not occur in both datasets. As we can see, the probabilities of the masdar patterns given by participants are a fairly close match to the probabilities of the same masdar patterns in the dictionary dataset. However, as noted, there are some patterns given by participants in the experiment that do not appear in the corpus (to the far left in Figure 4.13), and likewise, there are some patterns that are attested in the corpus that participants do not use (along the bottom of Figure 4.13). Both of these cases are on the lower end of the type frequency

spectrum, although there are non-occurring patterns in the experiment that have similar corpus frequency to patterns that do appear in the experiment, and patterns that do not appear in the corpus that have similar experimental frequency to patterns that do appear in the corpus. It is not entirely clear why speakers do not utilize the full range of masdar patterns, nor why they use masdar patterns from other verb forms, but the overall patterns remains that the log probability of a masdar pattern in the experiment is significantly correlated with the log probability in the corpus, when that probability is calculated over all verbs.

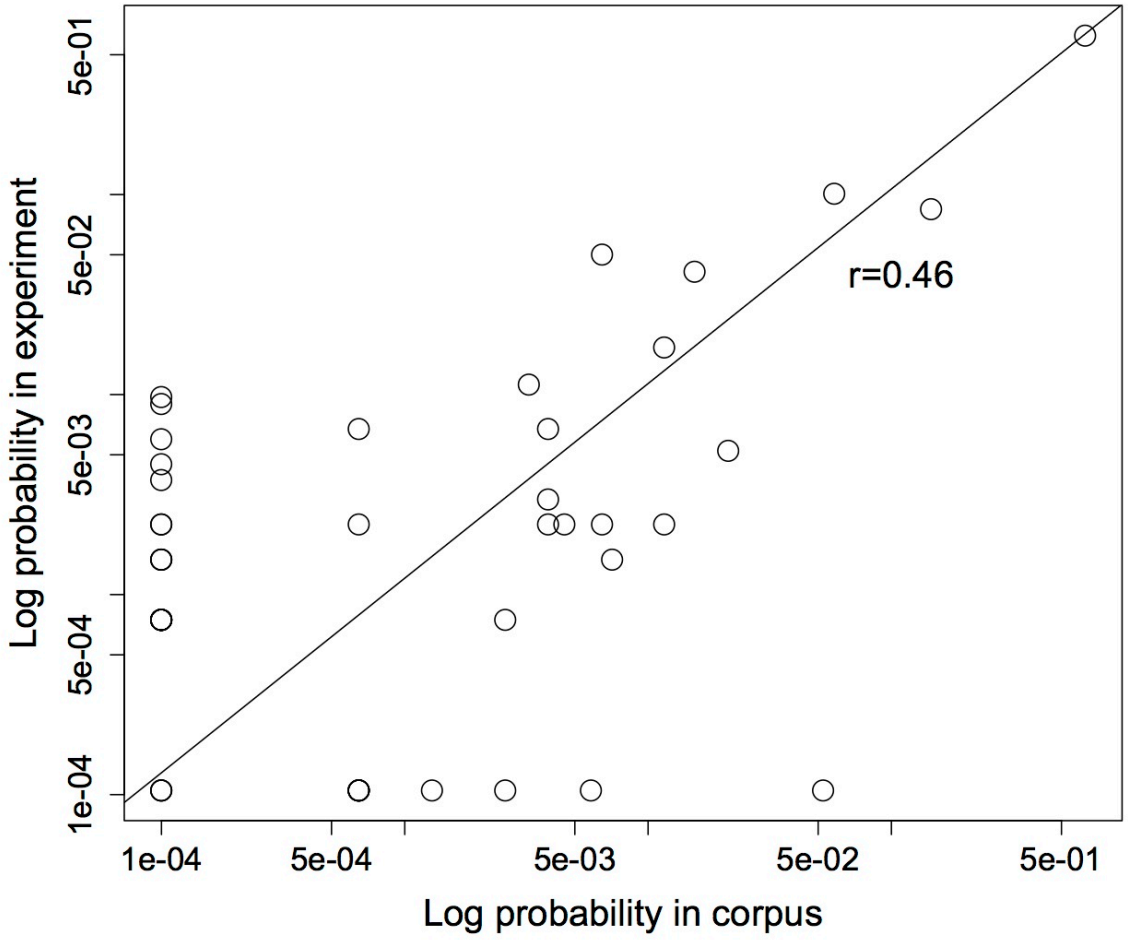

Figure 4.13: Expected vs. observed probabilities for masdar patterns

The major question under examination in this experiment is whether speakers know, and use, type statistics on the verb pattern in generalizing existing masdar patterns. Figure 4.14 shows the proportion of responses with each masdar pattern in the nonce responses, ordered by overall frequency across the verb patterns, such that all bars are in the same order in each of the three panels. The figure shows only the top ten response patterns, which captures the major differences among the three patterns.

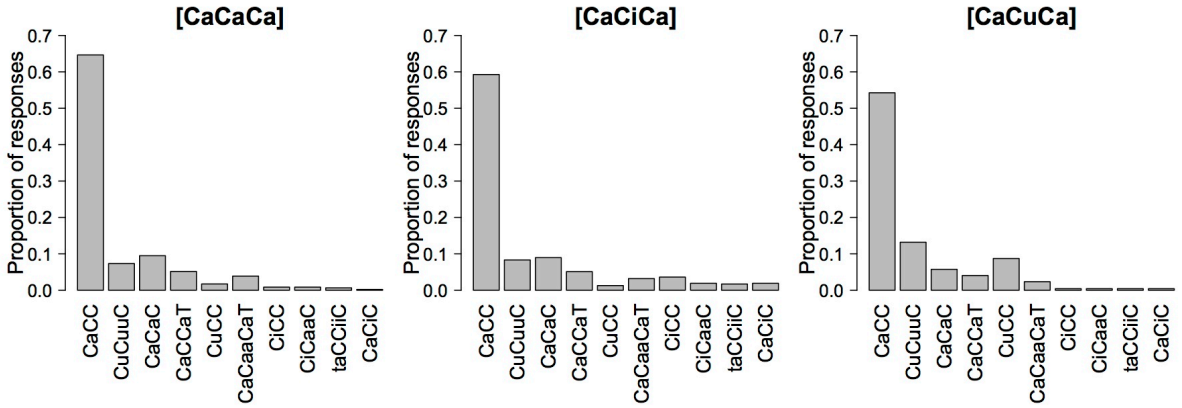

Figure 4.14: Masdar pattern responses by verb pattern, showing only top 10 patterns in overall responses

As is evident from Figure 4.14, there are few differences in the response patterns of participants across the three verb patterns. This is extremely surprising, as the modeling work in chapter 3 showed that type frequencies on the verb pattern are a very strong predictor of masdar pattern. On the basis of this, we would expect participants to primarily generalize the dominant masdar pattern for that verb pattern to nonce verbs, or to match the distribution of masdar patterns for that verb pattern (similar to what was observed in experiment 1A on the CV template). Rather, participants appear to use the same overall distribution of masdar patterns

regardless of the verb pattern of the nonce item. For all three verb patterns, [CaCC] is by far the most frequent response pattern. [CuCuuC] is the second or third most frequent response pattern for all verb patterns. One notable difference is the use of [CiCC] for [CaCiCa] verbs and [CuCC] for [CaCuCa] verbs, where those masdar patterns are extremely infrequent for the other verb patterns. This suggests that there may be some separable processing of the CV template and the vocalic melody, whereby the CV template of the masdar may be dictated by the CV template of the verb (which in this case is static), and the vocalic melody of the masdar is dictated by the vocalic melody of the verb. However, there are no other sets of masdar patterns that share the same CV template that differ in the vowels across the verb patterns in this same manner. One small but possible supporting piece of evidence for this theory is the generally more frequent use of [CuCuuC] for [CaCuCa] verbs, but the comparison patterns that have the same CV template, [CaCaaC] and [CiCaaC], do not seem to vary across verb patterns. Nonetheless, it is clearly the case from the experiential data that participants are not attending to the highly-predictive cue of the verb pattern, and seem to employ a similar strategy in generalization regardless of the verb pattern. The differences between verb patterns in the frequency of [CaCC], [CiCC], and [CuCC] suggest that there may be a small influence of the vowel quality of the verb on the vowel quality of the masdar pattern, but this is relatively weak evidence.

4.3.3.2 Filler results

As noted in the methodology, participant accuracy on filler items was both quite low and highly variable, with a mean accuracy of 59.7% (S.D.=14.3). Accuracy on individual filler items across participants was similar on average but even more variable, with a mean accuracy of

60.2% (S.D.=33.24). For one verb, [kasara], 100% of participants entered the correct masdar, while for other verbs, overall accuracy was as low as 5% across participants. The majority of verbs (n=22) had masdar accuracy over 50%, but accuracy was quite variable as noted. One possible reason for this variability is the token frequency of the masdars of the filler verbs. Figure 4.15 below shows the frequency of the masdar versus the accuracy across participants for each filler verb. Log frequency of the masdar is significantly correlated with accuracy, r=0.54, p<0.001. One major point to note is the wide disparity in accuracy for low-frequency filler items. For some filler items that are very low or zero frequency, accuracy is extremely low, while for others, accuracy is over 90%.

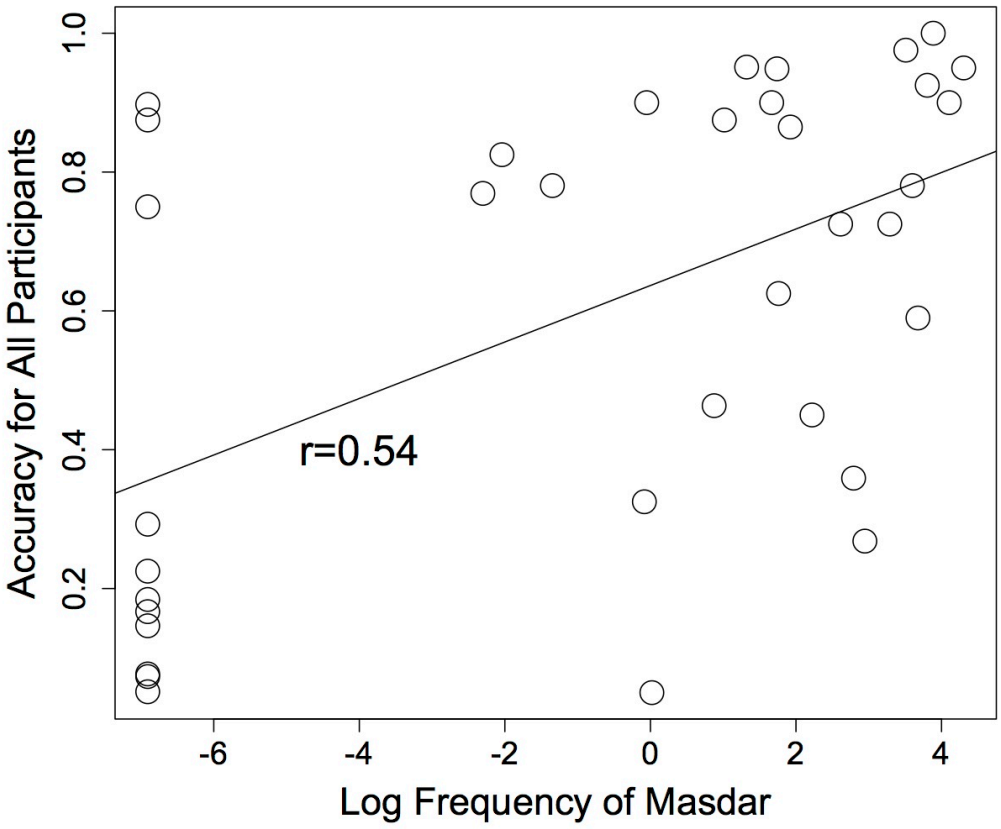

Figure 4.15: Log frequency of filler masdars vs. accuracy

A second possible source of the variance in accuracy among low-frequency items is that participants are more accurate on verbs that take a more frequent masdar pattern. Because verbs with the pattern [CaCaCa] are much more frequent overall than verbs with the other two patterns, it is possible that speakers will be more accurate on these filler verbs even if they are unfamiliar with them, as there is greater type support for the morphological pattern. The effect of masdar pattern frequency on participant accuracy was assessed using a mixed-effects regression model. By using this class of model, we can independently assess the effects of masdar pattern frequency and token frequency of the individual masdar, while also controlling for random effects of participant.

The model tried to predict whether a participant would choose the correct masdar for a given item using the fixed effects of log token frequency of the masdar, probability of the masdar pattern (from the corpus dataset), and an interaction between these two terms. In addition, random slopes for each of the fixed effects (except the interaction term) by participant were included. As before, I will assess which fixed effects are significant predictors of whether a participant will select the correct masdar used nested model comparison (Barr et al., 2013).

First, the log frequency of the masdar was a significant positive predictor of accuracy on the masdar, β=0.21, S.E.=0.02, $\chi^2(1)$=72.86, p<0.001. The probability of the masdar pattern was also significant positive predictor of accuracy on the masdar, β=3.52, S.E.=0.43, $\chi^2(1)$=43.63, p<0.001. The interaction between log frequency of the masdar and masdar pattern probability was not a significant predictor of accuracy, β=-0.13, S.E.=0.07, $\chi^2(1)$=3.12, p=0.07. We can take these results to mean two things. First, participants are generally more accurate on masdars for [CaCaCa] verbs (which have a higher masdar pattern probability) than on masdars for [CaCiCa] or [CaCuCa] verbs, and likewise, that participants are generally more accurate on masdars for

[CaCiCa] verbs than for [CaCuCa] verbs. Second, this indicates that, as noted above, the token frequency of the individual masdar is also predictive of accuracy. This confirms the correlation seen above in Figure 4.15. This indicates that both general pattern familiarity (via type frequency or pattern probability) and individual word frequency play a role in whether speakers are able to form the correct masdar for the filler verbs in this experiment.

4.3.3.3 Analysis of dialect background

4.3.3.3.1 Nonce items

As noted in the methodology, participants in this experiment came from a variety of different dialectal backgrounds. There are a number of phonological and phonotactic differences between dialects, which could influence the phonological or phonotactic shape of the masdars. In addition, it is common for different dialects of many languages to have different lexical items, stemming from language contact or other general mechanisms of language change. Although there appeared to be little variation in noun plural formation between participants from different dialect backgrounds in experiment 1A, there has been little, if any, examination of how Arabic dialects might vary in masdar formation. As noted, for the purposes of these analyses, participants were classified by major dialect region. 16 participants reported speaking a Levantine dialect, 9 participants reported speaking a North African dialect, 2 participants reported speaking a Peninsular dialect and 8 participants reported speaking an Egyptian dialect. For these analyses, only groups including at least 8 participants will be analyzed separately, although all participants will be included in the cross-dialect analyses.

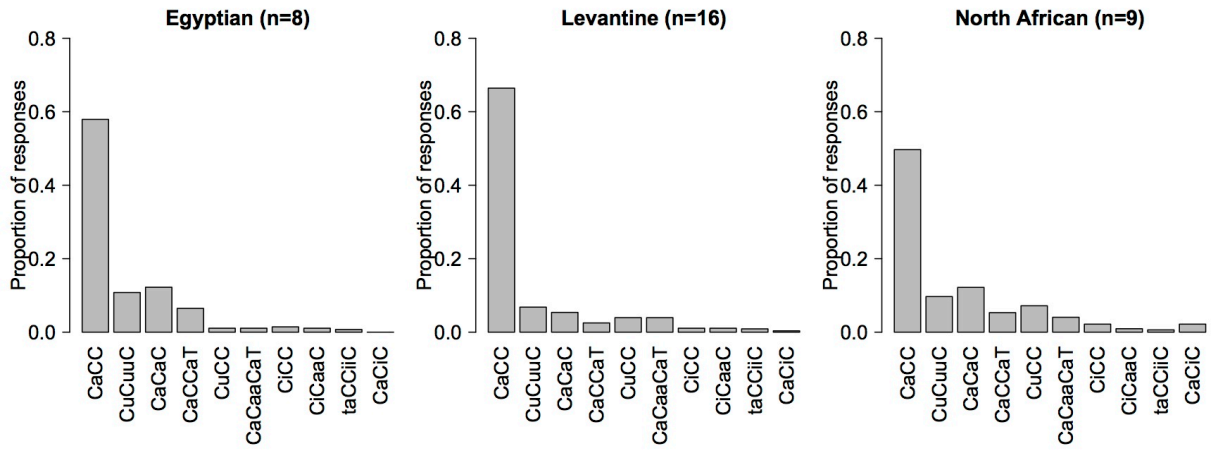

Figure 4.16: Masdar pattern responses by dialect, showing only top 10 patterns in overall responses

As shown in Figure 4.16, there seem to be some differences in how speakers of different dialect backgrounds generalize masdar patterns to nonce forms. In particular, the North African group shows the weakest preference for [CaCC], with 49.7% of responses having this pattern, while the Levantine group uses [CaCC] 66.4% of the time. The Egyptian group also shows a slight preference for [CuCuuC] relative to other dialect groups, using it for 10.8% of forms, where Egyptian speakers use it 9.7% of the time and Levantine speakers use it 6.8% of the time. The Levantine group also uses [CaCaC] slightly less than the other dialect groups, giving 5.4% of verbs this pattern, compared to 12.2% for the Egyptian group and 12.2% for the Levantine group.

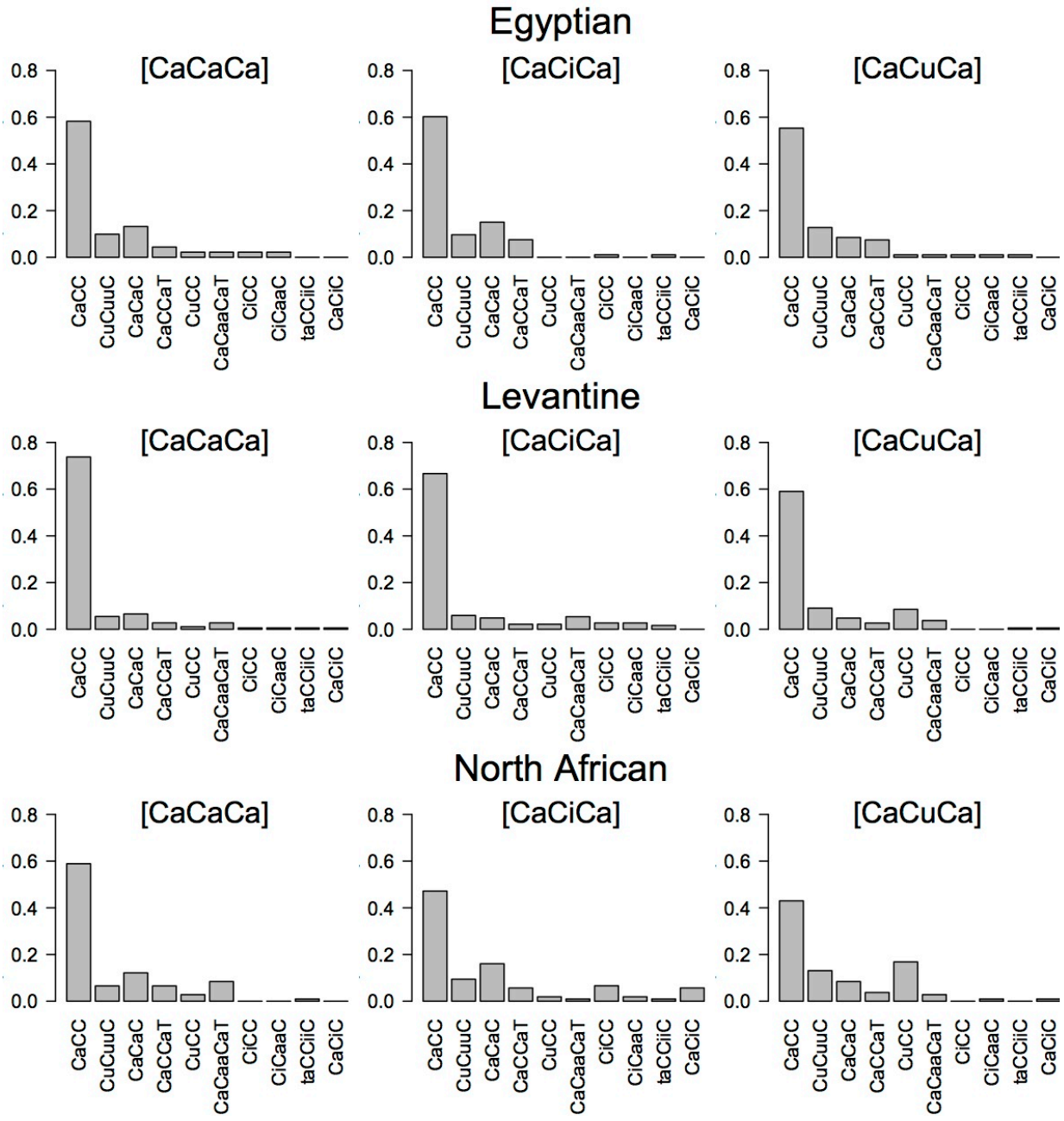

Figure 4.17: Masdar patterns by dialect by verb pattern, showing only top 10 patterns in overall responses

Figure 4.17 shows the same ordering of masdar patterns by dialect group, broken down by verb pattern. As with the main results, it is immediately apparent that no dialect group is using the verb pattern as a cue to masdar pattern in the way that would be expected from the

modeling results. For each dialect group, for each verb pattern, [CaCC] is the dominant response pattern. There are some minor differences among the dialect groups that are related to trends noted in the overall results. For instance, the North African group appears to be the driving group behind the differential use of [CiCC] and [CuCC] for [CaCiCa] and [CaCuCa] verbs, respectively, where the Levantine group shows smaller differences across the verb patterns in the use of these masdar patterns and the Egyptian group appears to show no difference at all. Nonetheless, it is quite clear that no dialect group capitalizes on the statistical patterns by verb pattern; if this were the case, we would expect to see strong differences in the use of [CaCC], [CaCaC], and [CaCaaCaT] across the verb patterns, but the experimental results show little, if any, evidence of differences.

As with the other experiments, the effect of dialect background on masdar generalization was also assessed using Krippendorff's alpha (Krippendorff, 1980). As noted previously, the alpha statistic computes the overall agreement above chance between raters on assigning *n* items to *c* categories, where 0 = chance agreement and 1 = perfect agreement. For this experiment, a 'rater' is a participant, an 'item' is a nonce verb, and a 'category' is a masdar pattern. Inter-speaker agreement was assessed for each dialect group, as well as across all dialect groups.

Across all dialect groups, α=0.0285. This indicates that participants overall are barely above chance agreement on the masdar pattern for a given input item. Within each dialect group, the agreement is also extremely low: Egyptian, α=-0.0299; North African, α=0.0148; Levantine, α=-0.0068. In total, the agreement within dialect group is actually somewhat worse than the agreement across dialects. However, the overall agreement, both across dialect groups and within dialect groups is extremely low, hovering right around chance agreement (which would be 0 by this measure).

In sum, for the nonce, items, there is limited evidence of an effect of dialect background on masdar generalization. The aggregate results in Figures 4.16 and 4.17 indicate that there may be some differences across dialects in the propensity to generalize the dominant pattern [CaCC]. However, when agreement within a dialect group is assessed on an item-by-item basis using Krippendorff's alpha, the agreement on the masdar pattern for a particular item is roughly equal across the dialect groups. The general tendency of Levantine versus North African speakers to use [CaCC] overall is interesting, and warrants further investigation, as it suggests that speakers may be either utilizing different strategies in generalization or drawing on differing statistical distributions of masdars in the differing dialects. Nonetheless, this pattern only appears when results are considered in aggregate, which suggests that speakers of different dialects do not treat individual nonce verbs in a consistent manner.

4.3.3.3.2 Filler items

The effect of dialect background on filler items was also assessed. As noted previously, there was a wide range of accuracy both across participants and across items. First, the general effect of dialect background on accuracy is shown in Figure 4.18. Although it appears on the basis of this graph that the North African group shows higher accuracy than the other two groups, this is not a statistically significant difference, $t(14.92)=2.03$, $p=0.61$.

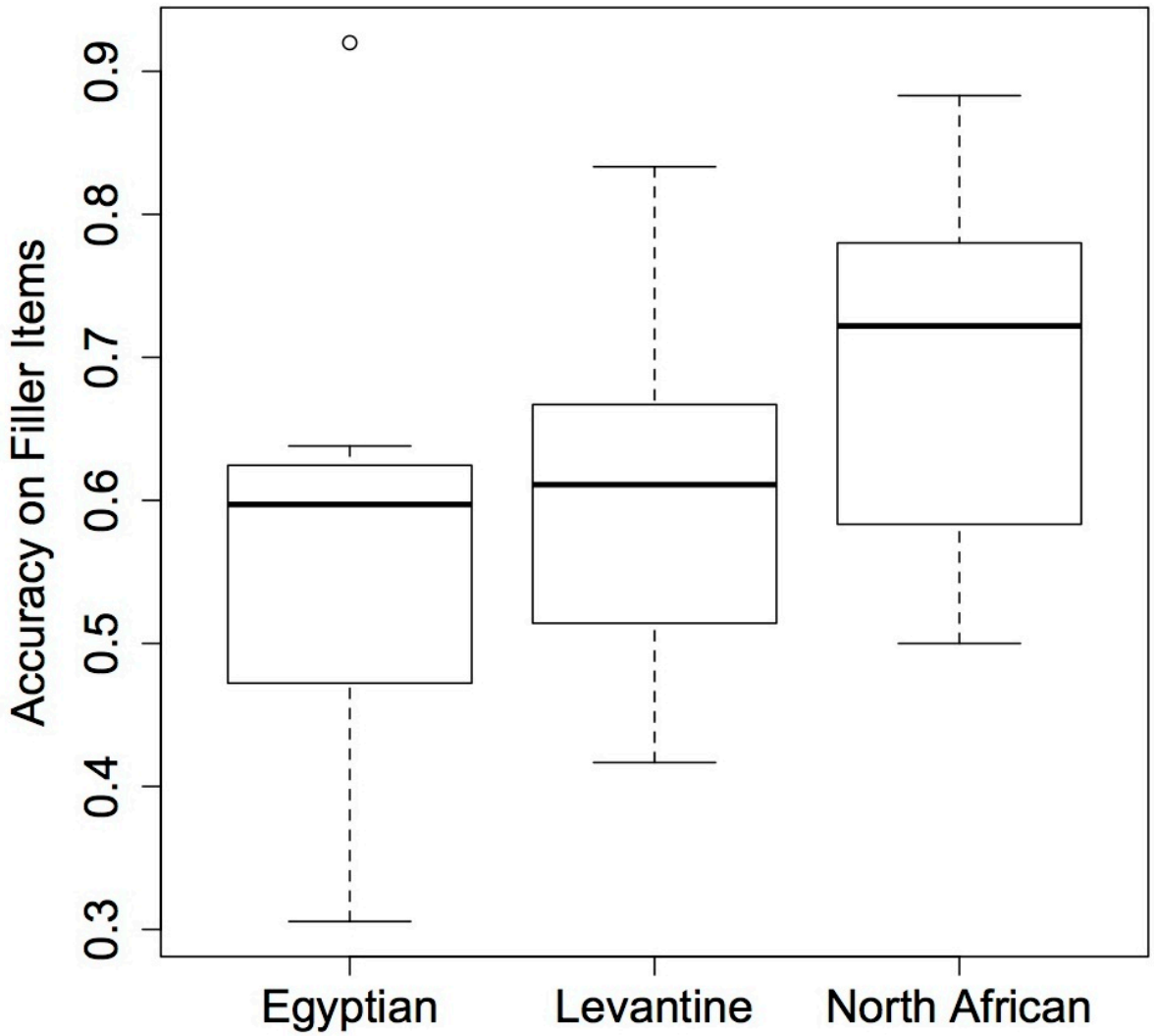

Figure 4.18: Boxplot of filler accuracy for three major dialect groups

For the filler items, as noted previously, there was a wide range in accuracy across items and individuals. Thus, another possible source of information on whether dialect groups consistently form masdars in a different manner is assessing the agreement on the masdar for filler items that an individual answered incorrectly. In experiment 2, the overall pattern of results for filler and test items suggested that speakers treat test items similarly to filler items when they

do not know the correct answer for the filler item. Thus, speakers may treat filler verbs for which they do not know the masdar similarly to nonce items.

In order to assess whether speakers of different dialect background were more consistent in forming masdars for filler items, I computed Krippendorff's alpha across all speakers as well as within dialect groups for filler items on which participants did not give the correct answer. Generally, then, this measure indicates the extent to which participants agreed on the masdar pattern when they did not give the correct masdar pattern for a filler item.

Across all dialect groups, $\alpha=0.134$. This indicates that participants overall are barely above chance agreement on the masdar pattern for a given filler item, when they were incorrect on that filler item. Within each dialect group, the agreement is also extremely low: Egyptian, $\alpha=0.127$; Levantine, $\alpha=0.17$; North African, $\alpha=0.101$. As with the nonce items, the agreement on masdar forms for individual items is very low across the board. Interestingly, it is slightly higher than the agreement on masdar forms for nonce verbs, but both are extremely low compared to the alpha statistics for the nonce noun plurals in experiment 1A, where across all participants, $\alpha=0.368$. In sum, for both the nonce and the filler items, we see little evidence of dialect background in forming masdars. Moreover, participants show extremely low agreement on the masdar for a particular nonce, item, which is another major point of difference between masdars and noun plurals. The discussion will delve into the possible reasons for the differences observed between the two systems in more detail.

4.3.3.4 Effects of frame sentence

One other possible source of variability across nonce items is the frame sentence in which each item was displayed. Although Wright (1988) suggested that the syntax and semantics of the verb play a role in masdar formation, I demonstrated in chapter 3 that there is little evidence for either of these factors having an effect on masdar form. Recall that the nonce items were randomized for sentence frame, in order to control for possible effects of the syntactic and semantic properties of the verb on masdar formation for nonce verbs in this experiment. Thus, participants saw the same nonce item in sentence frames with differing syntactic and semantic properties, which may have had an effect on masdar formation. In order to assess whether the sentence frame played a role in masdar formation, I again used Krippendorff's alpha (Krippendorff, 1980). However, unlike previous analyses using this statistic, an 'item' is the sentence frame rather than the nonce item. Thus, this statistic computes the overall agreement above chance of participants ('raters') on the masdar patterns ('categories') given the sentence frames ('items') in which the nonce forms were presented.

For all participants, α=0.0473. This indicates that participants overall are barely above chance agreement on the masdar pattern for any item with a particular sentence frame. This is similar to the overall agreement across participants on masdars for individual nonce items (α=0.0285). This indicates that the specific sentence frame in which a participant saw a nonce item had, at best, a tiny effect on the specific masdar pattern given for that item. Thus, based on both the evidence presented in chapter 3 and the analyses here, we can conclude that there is little, if any, evidence that the syntactic or semantic properties of the verb have an effect on the specific masdar form it will take.

4.3.3.5 Comparison to models of masdar formation

4.3.3.5.1 Model details

Although I demonstrated above that masdar frequencies in the lexicon are strongly correlated with masdar frequencies in the experimental data, this is only some evidence that speakers may be probability-matching to the frequencies of existing masdars in the lexicon when generalizing to nonce forms. Further, the general pattern of results indicates that speakers are not utilizing the verb pattern in generalizing existing masdar patterns to nonce verbs. To best determine whether speakers are using a probabilistic strategy, and whether they are using the verb pattern at all in generalization, I will compare four predictive analogical models to the experimental data. The approach taken here is similar to that in the model comparison in experiment 1A, but the factors examined are slightly different due to the different systems under examination.

The models used here are variations on the analogical models used in the model comparison in chapter 2. As a reminder, in an analogical framework, a test item (the nonce verb) is compared to a set of existing verbs (the comparison set), and the test item is predicted to take the masdar pattern of the most similar gang from the comparison set. Unlike in experiment 1A, none of the models use fine-grained segmental similarity in assessing similarity amongst forms. In the modeling work in chapter 3, there was no significant difference in accuracy between the model using type statistics on the pattern (*Simple Pattern Match*) and the one using these factors in addition to fine-grained segmental similarity (*Pattern-Restricted GCM*). In addition, the main question that remains to be answered is whether participants are using type statistics on the verb pattern or on the CV template of the verb. Thus, none of the models in the model comparison use

fine-grained segmental similarity in determining analogy. Rather, I vary the base of similarity between the CV template and the verb pattern.

I previously defined a morphological gang for the masdar as a set of verbs with the same verb pattern that also take the same masdar pattern. Under the assumption that speakers are using the verb pattern in forming masdars, this definition holds up well. However, if speakers are not tracking type statistics on the verb pattern, then this definition may be incorrect. If speakers are instead tracking statistics on the CV template, then the definition of a gang should also be different. Thus, for the models that assume the CV template as the base of similarity, the gangs are defined as a set of verbs that take the same masdar pattern (as all form I verbs have the same CV template). In contrast, for the models that assume the verb pattern as the base of similarity, the gangs are defined as a set of verbs sharing the same pattern that take the same masdar pattern. This differential definition of a gang is necessary for the models to accurately assess statistics by the relevant representation (CV template or pattern).

The definition of similarity is varied amongst the models as noted above, where in two of the models similarity is defined by a shared CV template, and in the other two models similarity is defined using only the CV Template (which for this system is the same as considering any form I verb in the lexicon), while in the other two models, similarity is defined on the level of the verb pattern. By comparing these two levels of abstraction as the basis of similarity, I can assess whether participants are at all using the verb pattern in forming masdars for nonce verbs. The decision rule used by the models is also varied, with two models using a deterministic choice rule and two using a probabilistic choice rule, in a 2x2 design. The specific parameters of the models and methods of implementation are discussed below.

Deterministic Template Match: This model uses type statistics on the overall masdar system, and uses a deterministic choice rule. Thus, in fact, for any given nonce verb, it will select the same output of [CaCC], as this is the most frequent pattern in the system. This model can be considered a baseline of sorts.

Probabilistic Template Match: This model also uses type statistics on the overall masdar system, but uses a probabilistic choice rule. The model considers all gangs in the system, and outputs a probability for that gang based on the number of items in the gang relative to the number of items in all gangs. This model thus generates a probability distribution over masdars for each nonce form.

Simple Pattern Match: This model uses type statistics on the verb pattern to determine the best gang with which to form an analogy, and selects from the candidate gangs deterministically. Thus, the model always selects the largest gang with a same verb pattern as the test item. Because this model is deterministic, the same gang will always be selected for a given singular, so the predicted masdars are generated from a single iteration of the model.

Probabilistic Pattern Match: This model also uses type statistics on the verb pattern, but uses a probabilistic choice rule. The model considers only gangs with the same verb pattern as the test item, and outputs a probability for that gang based on the number of items in the gang relative to the number of items in all gangs with the same verb pattern. As with the *Probabilistic Template Match*, the model generates a probability distribution over masdars for each nonce form.

4.3.3.5.2 Model fitting procedure

As with experiment 1A, I used the Jensen-Shannon (J-S) divergence (Cover & Thomas, 1991) to assess the fit between the experimental data and each model. As a reminder, the J-S divergence is a symmetric, weighted average of the Kullback-Leibler divergence (Kullback & Leibler, 1951), where the K-L divergence is the expected number of additional bits required to encode probability distribution Q using distribution P. Using the base 2 logarithm, $0<D_{JS}<1$. The closer the distribution of masdars between a model and the experimental data, the lower the divergence.

The distributions from all four models were first corrected using Laplace smoothing (Lidstone, 1920), such that any masdar pattern that was not predicted by the model was assigned a small not non-zero probability of 0.001, and likewise for any masdar pattern that was predicted by the model that did not appear in the experimental data. The divergence was then calculated between each model and the experimental results for each item across participants. The divergence was also calculated by participant by verb pattern across items. For each comparison (item or participant), the divergence was averaged for each model for each pattern, and for each model for each participant. These values were then averaged for each model and participant, for an aggregate value of how well each model fit the data overall. The best-fitting model is the one with the lowest divergence from the experimental data.

4.3.3.5.3 By-item fit

Table 4.1 below shows the mean divergence between each model and the experimental data by item. Probability distributions on the masdar patterns were calculated for each item

across participants, and for each model for each item. There were 180 nonce items, resulting in 720 total comparisons between the experimental data and the models. The divergence for each model was averaged across items by pattern (shown in rows labeled by pattern), and then averaged for each model across patterns (shown in row labeled "mean"). Recall that the lower the divergence, the better the fit between the distributions. The lowest divergence for each verb pattern and overall across patterns is marked in bold.

Table 4.1: Mean J-S divergence by item, by verb pattern

	Model name	Simple Temp Match	Prob. Temp Match	Simple Pattern Match	Prob. Pattern Match
	Level of similarity	CV Template	CV Template	Pattern	Pattern
	Decision rule	Deterministic	Probabilistic	Deterministic	Probabilistic
Pattern					
CaCaCa		0.2091	**0.1935**	0.2091	0.1970
CaCiCa		0.2486	**0.2245**	0.8071	0.6098
CaCuCa		0.2956	**0.2548**	0.9236	0.8756
MEAN		0.2511	**0.2242**	0.6466	0.5608

Overall, the best-fitting model is the *Probabilistic Template Match*, which uses the verb CV template to define similarity and a probabilistic choice rule. If we look at the individual patterns, we see sharp distinctions in fit between the models using the CV template as the basis of similarity (*Simple Template Match* and *Probabilistic Template Match*) and the models using the verb pattern as the basis of similarity (*Simple Pattern Match* and *Probabilistic Pattern Match*). For all three verb patterns, the CV template-based models show a better fit than the pattern-based models. Note, however, that the difference in fit between the CV template-based models and the pattern-based models is very slight for [CaCaCa] verbs, in particular relative to

the differences in fit for the other verb patterns. For this specific verb pattern, the most-frequent masdar pattern is [CaCC], accounting for about 80% of [CaCaCa] verb. The most-frequent masdar across all verb patterns is also [CaCC], accounting for about 65% of all verbs. Thus, the probability distributions for the two probabilistic models are very similar; and as noted, the probability distributions for the two deterministic models are the same. In contrast, for the other verbs patterns, the probability of the most frequent masdar by pattern and overall are quite different, leading to very distinct probability distributions between the template-based models and the pattern-based models. Thus, the [CaCiCa] and [CaCuCa] verbs are likely to be more informative in distinguishing between bases of similarity and decision rules.

For the decision rule, the two probabilistic models also show slightly lower mean divergences than the corresponding deterministic models. This pattern holds true for all three verb patterns, although note again that the differences for [CaCaCa] verbs are very slight relative to the differences for the other verbs patterns.

We can make a few interim conclusions about what these model fits mean. First, if we consider the entire system, the probabilistic model using the CV template is the best match, but this is only slightly better fitting than the deterministic model using the CV template. If we consider the basis of similarity (CV template vs. pattern), it is clear that overall, model fit calculated by item is much better when using statistics on the CV template than on the verb pattern. Thus, in general, for individual nonce verbs, participants in aggregate seem to be using a probabilistic strategy, and seem to be matching statistics on the CV template rather than the verb pattern.

4.3.3.5.4 By-participant fit

Table 4.2 below shows the mean divergence between each model and the experimental data by participant. Probability distributions were calculated for each participant for each verb pattern, and for each model for each verb pattern. There were 41 participants and three verb patterns, resulting in 492 comparisons between the experimental data and the models. The divergence was averaged for each model for each singular template (shown in rows labeled by patterns), and then averaged for each model across singular templates (shown in labeled "mean"). The lowest divergence for each verb pattern and overall across verb patterns is marked in bold.

Table 4.2: Mean J-S divergence by participant, by verb pattern

	Model name	Simple Temp Match	Prob. Temp Match	Simple Pattern Match	Prob. Pattern Match
	Level of similarity	CV Template	CV Template	Pattern	Pattern
	Decision rule	Deterministic	Probabilistic	Deterministic	Probabilistic
Pattern					
CaCaCa		0.2357	**0.2254**	0.2357	0.2266
CaCiCa		0.2649	**0.2416**	0.8147	0.6196
CaCuCa		0.3117	**0.2904**	0.9232	0.8819
MEAN		0.2707	**0.2525**	0.6578	0.5760

By participants overall, the best-fitting model is also the *Probabilistic Template Match*. Like the by-item fit, this model is also the best fit for all three verb patterns. As with the by-item fit, we see very large differences in fit across the three verb patterns. First, looking at the factor of level of similarity (CV template- or pattern-based), the fit is very similar for [CaCaCa] verbs for both types of models, while the fit for the other verb patterns is much better for the CV template-based models. A particularly interesting pattern that we see repeated in the by-item and

by-participant fit is the worsening of fit for [CaCiCa] verbs and [CaCuCa] verbs. If you recall, [CaCaCa] verbs are the most frequent across the system overall, accounting for about 75% of all verbs in the dataset, followed by [CaCiCa] verbs at about 18% and finally [CaCuCa] verbs at about 6%. Thus, we see an overall worsening of fit for the pattern-based models as the type frequency of the verb pattern decreases. This trend is also true for the CV template-based models, but the degree of worsening is much smaller.

Next, if we look at decision rule, overall, the probabilistic models fit better than their deterministic counterparts. The by-participant fits diverge somewhat from the by-item fits when considering the decision rule factor by pattern. For all three verb patterns, the probabilistic models both fit better than their deterministic counterparts. In sum, we see a similar general pattern of results by item and by participant, with the *Probabilistic Template Match* the best-fitting model overall, although by participant, we see slightly smaller differences across the verbs patterns in fit.

From the summed data above, it is difficult to tell if individual participants are using differing strategies in forming masdars. Table 4.3 shows the number of participants for whom each model fits the best of the four models. Overall, we see an interesting but important difference using this measure that using the summed data above, with 22 participants fitting the *Probabilistic Template Match* best and 19 participants fitting the *Simple Template Match* best. This shows that, while in aggregate, participants fit the *Probabilistic Template Match* slightly better than the deterministic *Simple Template Match*, individual participants show some variation, with only slightly more than half of the participants best fitting the probabilistic template-based model and slightly fewer than half best fitting the deterministic template-based model. This indicates that participants use somewhat different strategies in forming masdars

from nonce verbs, with some selecting masdars more deterministically (that is, selecting [CaCC] more often than would be expected and other masdar patterns less often than would be expected), and other participants selecting masdars more probabilistically. Critically, not a single participant fits best to either of the models that use the pattern as the basis of similarity, which indicates that participants in general are not utilizing the pattern as a factor in forming masdars. This is quite surprising given the robust effects of the pattern in the modeling work in chapter 3, and the possible explanations for this result will be considered in the discussion.

Table 4.3: Number of participants with best fit to each model

Model name	**Simple Temp Match**	**Prob. Temp Match**	**Simple Pattern Match**	**Prob. Pattern Match**
Level of similarity	CV Template	CV Template	Pattern	Pattern
Decision rule	Deterministic	Probabilistic	Deterministic	Probabilistic
Number of participants	19	**22**	0	0

The summed divergences by participant do not examine the amount by which a participant fits a particular model better, but only which model of the four fits best. Figure 4.19 shows the divergence from the *Simple Template Match* versus the divergence from the *Probabilistic Template Match* for each participant. Participants who fit the *STM* better are above the *x=y* line, while participants who fit the *PTM* better are below the *x=y* line. As with experiment 1A, we see that the difference in model fit is relatively small for an individual participant, although again the *STM*-fitting participants are closer to the *x=y* line than the *PTM*-fitting participants, which is in accord with the overall better for the *PTM* across participants. Similarly to experiment 1A, the divergences from the two models are significantly positively

correlated, r=0.80, p<0.001. This indicates, again, that the behavior of an individual participant is not entirely deterministic nor entirely probabilistic, but rather than individuals show a general tendency toward one decision rule.

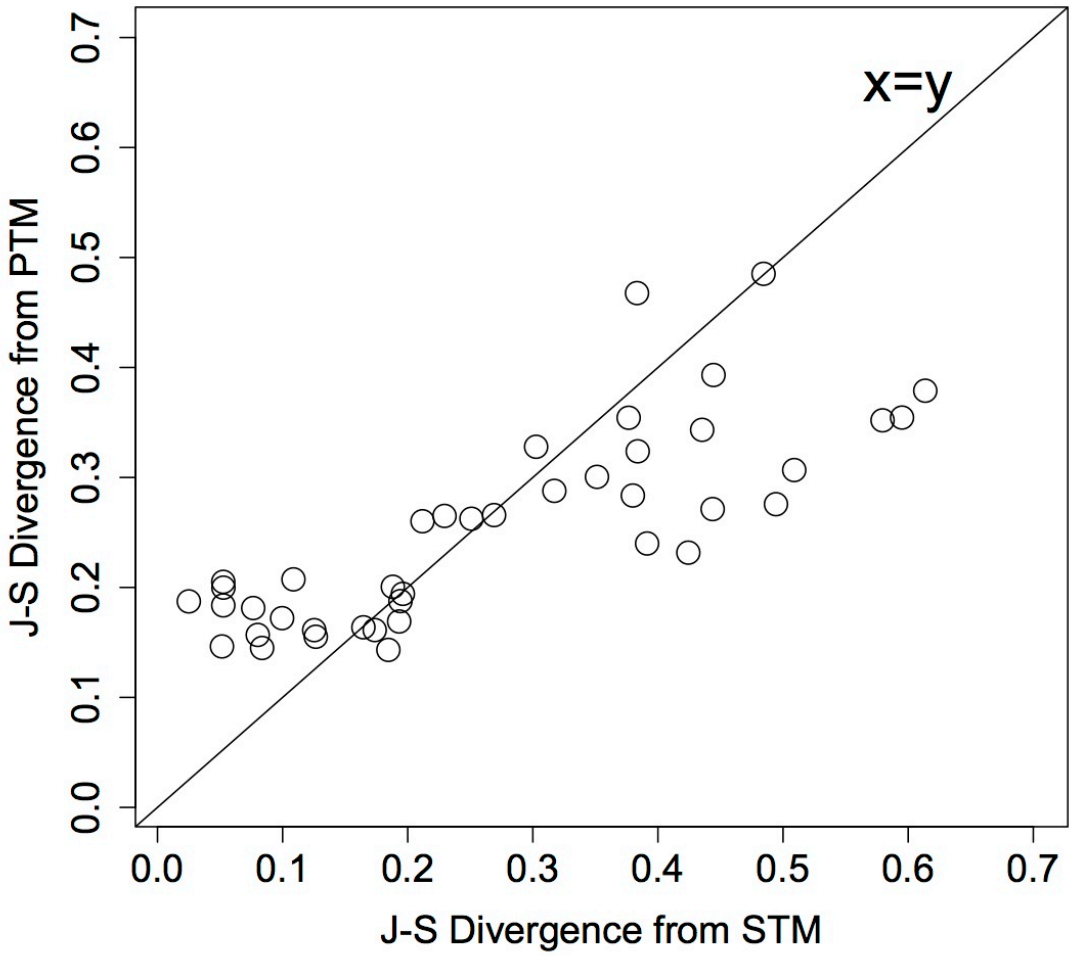

Figure 4.19: By-participant divergence from *Simple Template Match* vs. *Probabilistic Template Match*

As noted previously, participants showed a wide range of accuracy on filler items, with the lowest-performer achieving only 30.5% accuracy and the highest-performer achieving 92% accuracy. Accuracy on filler items could be considered a measure of general knowledge of the

masdar system, and thus we might expect to see some differences in how speakers generalize existing masdar patterns depending on their level of accuracy on the filler items. In particular, if a speaker achieves higher accuracy on the filler items, then they could be considered to have better access to the lexical statistics of the system, whereas if a speaker achieves lower accuracy, then they should have weaker statistics, since they know fewer word types. In generalizing existing masdar patterns to nonce forms, then, we would expect speakers who achieve higher accuracy on the filler items to show a more probabilistic strategy, as they should both know more masdar patterns and have stronger knowledge of the statistical distributions of those forms, whereas speakers who achieve lower accuracy on the filler items should show a more deterministic strategy. One way to assess this is to compare the filler accuracy of the two groups in Table 4.3 above. However, there is no significant difference in accuracy between participants who fit the *Probabilistic Template Match* than participants who best fit the deterministic *Simple Template Match*, $t(37.61)=1.85$, $p=0.07$. If we consider only the 8 participants who achieve at least 70% accuracy on filler items (the threshold used in experiment 2 for inclusion in analyses), we do find that seven of these participants fit the *Probabilistic Template Match* best. Nonetheless, in aggregate, there is not consistent evidence of an effect of accuracy on filler items on generalization strategy.

4.3.4 Discussion

Overall, this experiment demonstrates that speakers of Arabic are able to generalize existing masdar patterns in the language in a manner that reflects lexical statistics of the morphological system. Across speakers, we observe a general probability-matching in generalizing existing masdar patterns, which is reflected by the strong correlation between experimental and corpus probabilities of the masdar patterns. However, contrary to what the model results in chapter 3 would predict, speakers do not generalize masdar patterns using type statistics on the verb pattern, but rather seem to calculate masdar pattern probabilities on the CV template, or across the entire lexicon, which in this system are equivalent. This is surprising, as the three verb patterns display very distinct distributions of masdar patterns among existing verbs, and type statistics on the verb pattern predict over 80% of masdars for unseen forms in the analogical modeling presented in chapter 3. Thus, participants appear not to utilize this strongly-predictive cue of the verb pattern in generalizing masdar patterns to nonce verbs.

One conclusion we can draw from this data is that many speakers of Arabic do know the masdar system as a whole well enough to reproduce the lexical statistics in the system, even if they do not utilize the verb pattern as a cue to masdar pattern. However, as in experiment 2, participant accuracy on filler items was very low, averaging 59.7% (S.D.=14.3). This number is not directly comparable to experiment 1 or 2, as both had specific numeric thresholds of accuracy for excluding participants from analysis. For this experiment, the exclusion criterion was that participants had to have licit template structure and diacritization for at least 90% of items. The range of accuracy was very large, ranging from 30.5% to 92.0% accuracy for individual participants. Similarly to experiment 1A, we do find a split in individual decision rules in generalization, with slightly more than half of the participants fitting a probabilistic model

best and slightly fewer than half fitting a deterministic model best. As noted, this indicates a tendency toward one decision rule, not an absolute pattern in generalization behavior.

The finding that participants do not use the verb pattern in generalizing to nonce forms is surprising, but there are a number of possible reasons why this may be the case. First, the orthography may play a role, in that diacritics that mark short vowels are generally not written. One hypothesis is that speakers are used to reading without short vowels, and thus may ignore them even if they are available, as they were in all the experimental stimuli. However, there is evidence that full diacritization actually improves reading comprehension, even for high proficiency adult native speakers (Abu-Rabia, 1998, 2001, 2002). In fact, Abu-Rabia hypothesizes that speakers are particularly likely to attend to vowel diacritics when faced with an unfamiliar words. In the case of this experiment, then, where speakers are faced with completely unfamiliar nonce verbs, one would expect them to pay *more* attention to the voweling of the nonce verbs than the voweling of familiar words. Overall, the hypothesis that standard Arabic orthography is the reason that speakers do not utilize the verb patterns in generalizing masdar forms is unlikely given this evidence.

A second possibility is that speakers receive information on the vowel quality of the nonce verb in reading or speech but mutate or discard it. There is some evidence from the psycholinguistic literature that vowels are relatively more mutable in linguistic processes than consonants. van Ooijen (1996) used a task called "word reconstruction," in which speakers are given a pseudo word that differs from an existing word by one segment and asked to name the first real word that comes to mind. Participants were instructed to change either a vowel, a consonant, or given no instruction as to which to change ("sound change") to find the existing word. Overall, participants were faster to respond in the vowel change condition than the

consonant change condition. Additionally, in the sound change condition, participants were more likely to respond with a word that differed by a vowel segment than by a consonantal segment. An experiment using the same task in three different languages found the same pattern of results, suggesting that the findings were not due to differences in consonant to vowel ratios in phonemic inventories or the relative ease of substituting a vowel versus a consonant (Cutler, Sebastian-Galles, Soler-Vilageliu, & van Ooijen, 2000). Thus, it is possible that speakers in this experiment were also willing to "mutate" the vowels of the nonce verbs, most likely to the most-frequent verb pattern [CaCaCa]. However, the statistics on masdars for [CaCaCa] verbs in the corpus dataset are quite different from those observed in the experiment, with participants using the non-dominant patterns, especially [CuCuuC], [CaCaC] and [CaCaaCaT] much more frequently than would be expected if participants were matching statistics on [CaCaCa] verbs. The evidence from word reconstruction tasks would only be directly relevant if speakers in this experiment were not mutating the vowel, but rather maintaining uncertainty about its identity. Testing whether participants are definitively changing the vowel quality of the nonce verb versus maintaining uncertainty about its identity would require experimentation that is beyond the scope of this dissertation. There are, however, other reasons to suspect that speakers simple do not form strong representations of the vowel patterns for existing verbs. As noted previously, form I verbs are the only verb form in which the voweling pattern is variable (see Table 3.1). Thus, this is the only verb form for which the voweling pattern is not entirely predictable (although it is linked to aspect and transitivity, as demonstrated in chapter 3). In addition, some verbs in fact can occur with multiple vowel patterns; although this is a minority of verbs overall (n=20, or 1.2% of all single-listing verbs), this may nonetheless affect the certainty of a speaker about the vowel pattern for a particular verb, which in turn would affect the strength of the lexical representation.

As I have noted previously, Arabic corpus resources are limited by the nature of the orthography, and thus estimating the extent to which a particular verb surfaces with a particular voweling pattern is difficult. If speakers are unable to fully encode the vowel pattern for form I verbs, however, then they would be unable to form distinct representations of the statistics for a given verb pattern, which would lead to the pattern of results observed in experiment 3.

This brings me to a final possibility, which is that participants are in fact matching distributional statistics on the CV template, and not on the pattern. As noted previously, in other morphological systems in Arabic, including the noun plural system, the CV template is the most important factor in predicting the output form of the morphological process (e.g., Dawdy-Hesterberg & Pierrehumbert, 2014; McCarthy & Prince, 1990a). For the noun plural, the CV template of the singular is the primary predictor of the CV template of the plural, and the vocalic melody of the singular is the primary predictor of the vocalic melody of the plural (McCarthy & Prince, 1990a; Ratcliffe, 1998). Thus, in the noun plural system, we see some dissociation between the CV template and the specific segmental features that populate it for a specific word form. As noted, there is a small piece of evidence in this experiment that speakers may be using a similar process: the over-use of [CiCC] for [CaCiCa] verbs and [CuCC] for [CaCuCa] verbs relative to [CaCaCa] verbs. Although this is far from definitive evidence, this suggests that speakers may be processing the vowel separately from the CV template. Thus, in fact, speakers may be using a similar strategy in generalizing masdars to nonce verbs as they do in generalizing plurals to nonce singular nouns, despite the differences in the systems. Moreover, the verb forms (of which there are 10 generally-recognized forms) are largely distinguishable from each other on the basis of the CV template, with only one triad of forms that have the same CV template

(forms VII, VIII and IX)[13]. In addition, form I verbs are the only subset of the paradigm where the masdar varies substantially across verbs. Across the entire verb form paradigm, speakers can rely on the CV template in order to predict the masdar, with the exception of form I verbs. As such, it is possible that speakers do not learn to differentially rely on the pattern only for the subset of verbs where it is beneficial to do so. However, as noted, the CV template is identical for all verbs of this form. Thus, on the basis of this experiment alone, we cannot determine whether the CV template is active in generalization, or whether speakers are simply matching statistics on the set of existing form I verbs. Further experimentation is necessary to determine the extent to which the CV template can be implicated in this finding. Given the relevance of the CV template in other morphological systems in Arabic, however, this is a compelling conjecture to explain the observed results in experiment 3.

As with experiment 2, this experiment also brings up questions of why speakers of Arabic show a generally weaker knowledge of existing masdars in comparison to noun plurals. In experiment 1A, which used the same open response paradigm, only 13 participants were excluded for achieving less than 80% accuracy on filler items, while 61 achieved this threshold of accuracy. In comparison, in experiment 3, if we were to apply this same threshold of accuracy, 37 of the 41 participants would have to be excluded. As discussed in experiment 2, even though there was no significant difference in frequency between the noun plural fillers and the masdar fillers, participants across the board achieve much lower accuracy on existing masdars. I will

[13] Form VII [ʔinfaʕala], form VIII [ʔiftaʕala], and form IX [ʔifʕalla] do share the same CV template for the masdar, which are [ʔinfʕaal] and [ʔiftiʕaal], respectively. However, form IX is quite rare and semantically restricted to verbs related to color . In addition, forms VII and VIII share the same CV template in the verb in addition to the masdar. Thus, this overlap does not invalidate the conjecture that the CV template is the most relevant means of predicting the masdar across the verbal paradigm.

come back to this issue of the general learnability of the noun plurals and masdars in the general discussion, and examine possible reasons why we see such large differences in knowledge of the two systems.

4.4 General discussion

The two experiments in this chapter demonstrate some interesting conclusions about the masdar system of form I verbs. First, experiment 2 demonstrates that many verbs that purportedly have multiple masdars according to an authoritative English-Arabic dictionary (Wehr, 1976) are undergoing or have undergone leveling to a single dominant masdar. The propensity of a speaker to select a particular masdar for a given verb is influenced primarily by the token frequencies of the two masdars. We find similar results for the filler items, where the likelihood of selecting the correct masdar is influenced primarily by the token frequencies of the actual masdar and the distractor masdar. Importantly, we do not find large differences in masdar preference for the multiple-listing verbs for speakers of different dialects, which rules out one of the major hypothesized sources of the verbs having multiple masdars in the first place. A second major hypothesis for the existence of multiple masdars is that the different masdar indicate different meanings of the verb. However, in experiment 2, the sentence frames were held constant for each item across participants, so the observed effects cannot be attributed to syntactic or semantic aspects of the verb or sentence frame. The possibility that the semantics differ for the two masdars has not been ruled out entirely, but given the lack of effect of semantics on masdar form in the analyses in chapter 3 and in experiment 3, this seems an unlikely explanation. In sum, experiment 2 demonstrated that, in general, speakers show about

80% agreement on the masdar for verbs that are purported to have two masdars, and the choice between the masdars is largely explained by token frequencies of the masdars. Thus, many verbs that reportedly have multiple masdars likely do not have multiple active masdars in the language, or are undergoing leveling to the extent that speakers show a general preference for one form over the other. Thus, the prevalence of multiple masdars for form I verbs is likely overstated in the Wehr dictionary, and in the literature in general.

Experiment 3 demonstrates that Arabic speakers, on the whole, generalize existing masdar patterns in a manner that reflect the type statistics of the masdar system of form I verbs. However, contrary to the predictions of the analogical model comparison in chapter 3, speakers do not seem to utilize the verb pattern in generalization to nonce verbs, but rather match statistics on the CV template, or on the entire form I verb system. In a comparison to four analogical models that vary in the level of similarity (CV template or pattern) and the decision rule (deterministic or probabilistic), we find that by item and by participant, the experimental data overall matches the model which uses a probabilistic decision rule on the CV template in determining the best masdar pattern for an unseen verb. As noted, on the basis of experiment 3 alone, we cannot determine whether speakers are probability-matching statistics on the CV template or statistics on all form I verbs. However, the conjecture that speakers are matching statistics on the CV template is compelling, given the importance of the CV template in other morphological systems in Arabic. If this is the case, then speakers may be utilizing a more uniform strategy in forming morphophonological representations across different morphological subsystems than previously thought.

In experiment 3, we see that pattern probability of the masdar is an important force in creating the masdar for an unseen verb. This is in line with a large body of literature in

morphology showing that type frequency is a driving factor in generalization (e.g., Albright, 2002; Alegre & Gordon, 1999; Daelemans et al., 1994; Ernestus & Baayen, 2003; Rumelhart & McClelland, 1986; Stemberger & MacWhinney, 1988). In addition, we find that speaker uncertainty about the optimal morphological pattern to apply to a word, generally, does not lead to deterministic behavior. In experiment 2, uncertainty is modulated by relative familiarity with the two masdar options as indicated by token frequency, but we nonetheless see cases of verbs where speakers are split nearly evenly on which form they prefer. In experiment 3, uncertainty a primarily a factor of the relative probabilities of the masdar patterns, but also of the large number of possible patterns, and we see probability-matching behavior in the face of these two sources of uncertainty. This probability-matching behavior in the face of inconsistency in the probabilities of possible outcomes has been demonstrated across a number of domains (Ernestus & Baayen, 2003; Hayes et al., 2009; Hudson Kam & Newport, 2005; Walter, 2011), but as mentioned previously, little literature has demonstrated probability-matching when there are a large number of possible patterns (c.f. Walter, 2011). The relative importance of these two factors (uncertainty due to inconsistency in relative probability and uncertainty due to many possible outcomes) in driving probability-matching behavior remains to be disentangled.

Further, we find an interesting split in participant strategy in generalization, with slightly more than half of the speakers better fitting the deterministic version of the model and slightly fewer than half of the speakers better fitting the probabilistic version of the model. Thus, individual differences in generalization strategy play a role in these results. As noted, there is a similar split in generalization strategy by participants in experiment 1A, where roughly half fit a deterministic model best, and roughly half fit a probabilistic model best. To my knowledge, this type of individual variation in generalization strategy has not been demonstrated for native

speakers in their L1. This does not seem to be tied to knowledge of the masdar system, as assessed by accuracy on the filler items. This is in line with Schumacher et al. (2014) and Hudson Kam and Newport (2005), who report individual differences in the propensity to over-regularize (which mirrors the deterministic strategy in this work) versus probability-match in artificial language learning tasks. The novel contribution of the current study is that this trend holds true even for native speakers of a language in their L1. Further experimentation is necessary to determine the source of these differences in generalization strategies.

One critical issue that remains unanswered is that participant accuracy on the filler items in the masdar experiments (2 and 3) shows that native speakers of Arabic do not know the masdar system as well as would be expected given both the predictability of the system as a whole and speakers' ability to reproduce masdar type statistics in generalization to new forms. In particular, the comparison to participant accuracy on the noun plurals in experiments 1A and 1B is a very intriguing and relevant one, given that noun plurals are (relatively) less predictable than the masdars, with the most-predictive model of the noun plural system achieving about 66% in analogical modeling (Dawdy-Hesterberg & Pierrehumbert, 2014), compared to about 80% for the masdar system (this thesis, chapter 3). However, participants in these experiments achieved lower accuracy on masdars for existing verbs than they did on plurals for existing nouns. In fact, native speakers achieved lower accuracy on the masdars in experiment 2 than they did on the noun plurals in either experiment 1A or 1B, despite 1A being an open response task. This is surprising for multiple reasons. First, as demonstrated by the analogical modeling on the masdar system, the masdar system as a whole is more predictable than the noun plural system, with similar models achieving 83% accuracy for masdars (this thesis, chapter 3) and 66% accuracy for noun plurals (Dawdy-Hesterberg & Pierrehumbert, 2014). Second, a forced-choice task, which

was used in experiment 2, should by default have higher accuracy than an open-response task, which was used in experiment 1A. The baseline for a forced-choice task with two options is 50%, while the baseline for an open response task is 0%.

Because the thresholds of accuracy for inclusion in the experiments were slightly different, the full results cannot be compared directly. However, to give one comparable measure for these two experiments, we can limit the sample to participants in experiment 2 that achieved the higher threshold of accuracy used in experiments 1A and 1B of 80%, and compare this to accuracy in experiments 1A and 1B. For experiment 2, the mean accuracy for participants achieving at least 80% accuracy[14] on fillers items was 85.4%, whereas for experiment 1B, the forced-choice plural experiment, mean accuracy was 94.1%, which is significantly higher, $t(44.74)=8.307$, $p<0.001$. Participants were also slightly more accurate in the open-response experiment 1A, achieving a mean accuracy of 87.4%, than in the forced-choice experiment 2, although this difference is not statistically significant $t(44.40)=1.895$, $p=0.65$.

There are a few reasons why this might be the case. First, it is possible that the masdars used as filler items in these experiments are less frequent than the plurals used as filler items. In the modeling work in chapter 3, I assumed a lower bound of token frequency >0 in Aralex, such that any word that appeared at least once was included in the set. However, the lower the frequency is for a given item, the less likely speakers are to know it. If token frequency is lower for the filler masdars in experiment 2 than for the filler plurals in experiments 1A and 1B, then this would explain at least to some extent the lower accuracy, as speakers are less likely to know the specific masdars that are tested in the experiment. If we examine the frequency of the filler

14. Note that the actual threshold used for inclusion in experiment 2 was 70% accuracy on filler items, as noted in the methodology. The threshold of 80% accuracy is used only for this statistic in order to compare accuracy directly to that in experiment 1B.

plurals from experiments 1A and 1B versus the frequency of the filler masdars from experiment 2, we find that there is no significant difference, t(73.78)=0.69, p=0.49. That is, there is significant overlap in the distributions of frequencies for these two sets, as shown in Figure 4.20. As noted in the methodology, there are some filler masdars that do not appear in Aralex (which appear on the far left of Figure 4.20), but we nonetheless do not find a significant difference between the frequencies of the masdar and noun plural fillers on the whole.

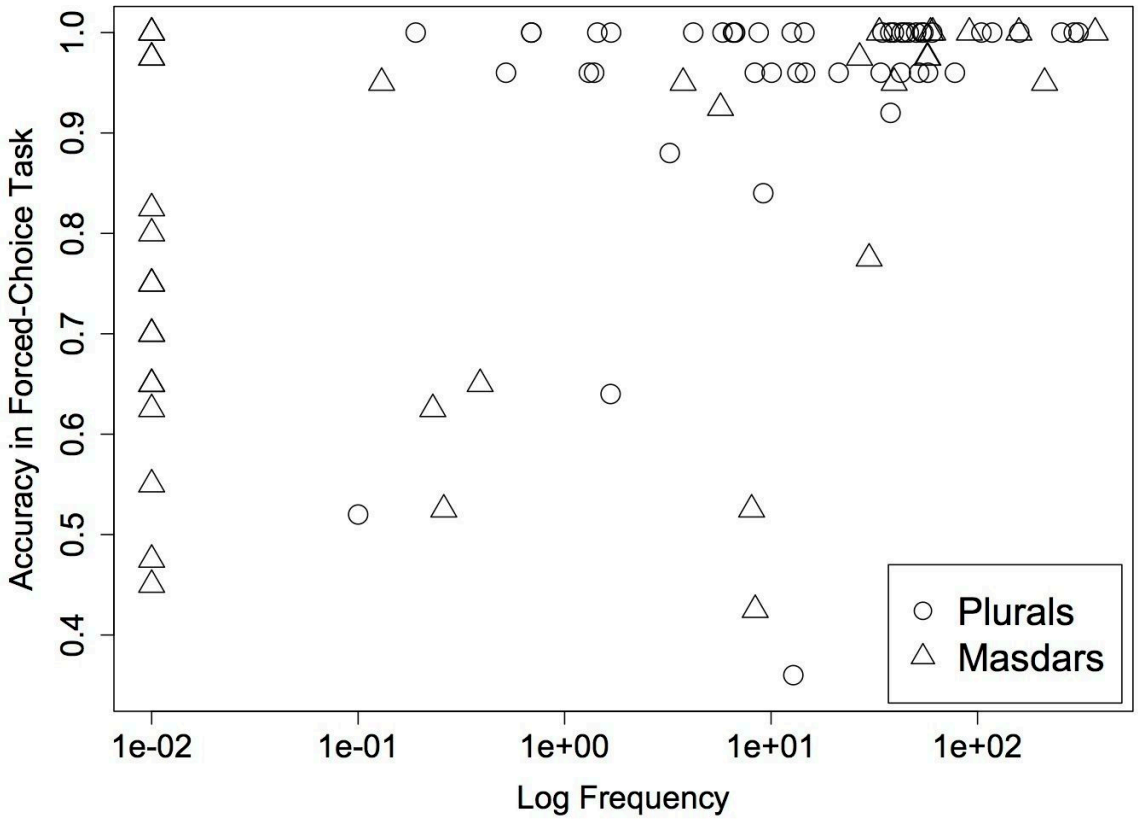

Figure 4.20: Log frequency vs. accuracy for experiments 1B and 2

If frequency based on the estimates from Aralex does not explain the difference in accuracy between the noun plurals and the masdars, then there are a few other possible reasons why speakers may be less accurate on the filler masdars. First, it is possible that the frequency of

the plurals and masdars is different in everyday use than in the type of text in the Aralex corpus (Boudelaa & Marslen-Wilson, 2010), which is largely composed of newswire text. It is well known that written Modern Standard Arabic is a distinct dialect from the spoken dialects used in everyday life, and so if masdars are more often used in formal registers, then it is possible that the corpus over-represents the frequencies of the masdars relative to what speakers encounter in normal life (see also Holes, 2004). However, it is difficult to verify this hypothesis, as spoken corpora tend to be much smaller, and transcribed Arabic corpora of dialectal speech are rare. In a similar vein, it is also possible that noun plurals are underrepresented in the Aralex corpus. In either case, there may be a true difference in frequencies between the noun plurals and the masdars that is undetectable given the corpus resources at hand. Unfortunately, this is a question that cannot be definitively answered at this time.

Chapter 5 : Conclusions and future directions

5.1 Overall discussion

The goal of this thesis is to investigate learnability and generalization of the morphology of Modern Standard Arabic, focusing on the relatively well-studied noun plural system and the understudied masdar system of form I verbs. As described in the introduction, the primary question under investigation is how speakers of Arabic form new words based on the structural characteristics and statistical distributions of existing words in the morphological system. The learnability of a system is a critical element in generalizability, as a speaker must know the existing words and patterns in the system well enough to extend them to new forms. There are two main elements of linguistic complexity that influence the learnability of a system. First, systems in which the complexity of changes between the input and output forms is greater should be harder to learn, as speakers must generalize across a wide range of words in order to capture complex changes such as those in non-concatenative morphology. Second, systems with higher regularity of correspondence between input and output forms should be easier to learn, as the output for a given input is more predictable.

Speaker behavior in generalization is a window into what speakers know about the linguistic system under investigation. In generalization, we can see the influence of these two factors of complexity of learnability. First, how speakers form analogies to existing words allows us to determine what kind of generalizations they have made about the input-output correspondences in the system. Second, speakers select more probabilistically or deterministically among possible outputs depending on the amount of uncertainty (or predictability) in the system.

The learnability of the noun plural system is fairly well established, and the existing theoretical and computational literature shows that speakers should be able to learn the system primarily on the basis of type statistics on the CV template (Dawdy-Hesterberg & Pierrehumbert, 2014; McCarthy & Prince, 1990a). Fine-grained segmental similarity to existing forms provides a small additional amount of predictive power in analogical modeling. In generalization, then, speakers have two choices for the base of analogy: the CV template, or the CV template plus additional shared segmental features. Speakers must also among the possible output choices, and could select probabilistically or deterministically.

In chapter 2, Experiments 1A and 1B demonstrated that native speakers of Arabic are able to match statistics in the lexicon in generalizing existing noun plurals to new forms. Across these two experiments, we find that speakers draw on both coarse-grained generalization across word types (the CV template) and fine-grained segmental information in determining similarity to existing forms. Both of these sources of information play a role in the likelihood of a speaker to generalize an existing plural pattern to an unseen singular in an open-response task, as well as to select among possible plurals in a forced-choice task. The relative size of the effects of these two sources of information (course- and fine-grained) is quite different than in other languages; in Arabic, the coarse-grained CV template provides the bulk of the predictive power, with fine-grained segmental features adding some additional predictive power, whereas in concatenative languages like English and Dutch, the relative importance of fine-grained segmental features in analogy-formation is much larger (e.g., Alegre & Gordon, 1999; Ernestus & Baayen, 2003). Type statistics on existing words play an important role in the process of analogy formation, with speakers generalizing plurals in a probabilistic manner that correlates with the probabilities of the plural templates in the lexicon. Critically, these type statistics are drawn at the level of the

CV template. In aggregate, speakers tend toward probability-matching over deterministic behavior even in a system with a large number of possible outcomes and a high degree of uncertainty as to the optimal plural for a given singular. Additionally, we find further evidence for individual differences in generalization strategy, with roughly half of the speakers in experiment 1A utilizing a more probabilistic strategy, and roughly half utilizing a more deterministic strategy. The latter group does not show entirely deterministic behavior; rather, they tend toward over-regularization of the dominant plural patterns and under-regularization of the less-frequent plural patterns.

The learnability of the masdar of form I verbs has not been established in the literature, with many sources citing it as unpredictable (Grenat, 1996; Holes, 2004; Kremers, 2012; McCarthy, 1985; Ryding, 2006). The first part of the analysis of this system was to determine how learnable the masdar system is, based on the two factors mentioned above.

The computational analyses in chapter 3 demonstrated that the masdar of form I verbs is predictable using type statistics on the verb pattern. Although classic grammars have indicated that phonological, syntactic, and semantic features all play a role in masdar formation (Wright, 1988), this chapter demonstrates that the phonological representation of the verb pattern, in conjunction with type statistics on existing forms, predicts over 80% of masdar patterns for existing verbs. There is little evidence that the syntactic features of the verb play an independent role in masdar formation, as they are strongly correlated with the verb pattern. In addition, we find no evidence for the influence of semantic features on masdar formation. Thus, the masdar for an unseen verb is best predicted in an analogical model using type statistics on a coarse-grained phonological generalization, namely the verb pattern.

Experiments 2 and 3 examined two facets of the masdar system. Experiment 2 examined verbs that purportedly have two existing masdars, and found that speakers agree on preferred masdar for a given verb about 80% of the time. The specific choice of masdar pattern for these multiple-listing verbs is statistically predicted primarily by the token frequencies of the two masdars. This indicates that many of the verbs in this experiment are undergoing, or have undergone leveling. Overall, the prevalence of verbs with multiple masdars forms is likely overstated, and dictionaries should be updated accordingly. Experiment 3 examined generalization of existing masdar patterns to nonce verbs, and found that while speakers overall generalize masdar patterns probabilistically, they do so not on the basis of statistics on the verb pattern as would be predicted by the modeling work in chapter 2, but rather on statistics across the range of existing form I verbs, or potentially on the CV template. We also find individual differences in the propensity to use a deterministic versus a probabilistic decision rule in generalization, the source of which is unclear.

Despite many differences between the noun plural and masdar systems in Arabic, we overall see similar tendencies in generalization. First, in both systems, speakers use a coarse-grained generalization as the basis of similarity in analogy formation. For the noun plural system, this generalization is the CV template, and we also see some evidence that speakers use additional fine-grained segmental similarity in analogy formation. For the masdar system, we also see some evidence that speakers form analogies on the CV template, and that additional fine-grained segmental similarity may play a small role in this process via the small differences in masdars across verb patterns. Second, in both systems, in aggregate, speakers best fit models which use a probability-matching strategy in generalization. Further, in both systems, we find a similar pattern of individual differences in generalization strategy, with roughly half of the

speakers in experiments 1A and 3 fitting a probabilistic model best, and roughly half fitting a deterministic model best.

These results provide interesting insights into the learnability of these two systems. First, for the noun plural system, the experimental results show that speakers generally use the base of similarity in generalization predicted by the theoretical and modeling work (Dawdy-Hesterberg & Pierrehumbert, 2014; McCarthy & Prince, 1990a). Thus, speakers are able to learn this system by forming coarse-grained generalizations on the level of the CV template, while also maintaining word-specific segmental information. Speakers then use both of these sources of information in generalization to unseen forms. Moreover, speakers match probabilities of existing words on the level of the CV template, which indicates that they track lexical statistics on this abstract generalization. Second, for the masdar system, the experimental results conflict with the modeling results, in that speakers do not seem to use the verb pattern as the base of similarity in generalization. This suggests that speakers do not form generalizations on the level of the verb pattern, despite it being a highly-predictive cue in analogical modeling. Rather, one strong possibility is that speakers form analogies on the basis of the coarse-grained CV template, although as noted, further experimentation is necessary to examine this conjecture. In addition, we see some evidence that speakers also use fine-grained segmental information in analogy formation via the slight differences across the verb patterns.

Like the noun plural system, the decision rule used by speakers for the masdar system is in aggregate probabilistic. However, this varies across speakers, with slightly fewer than half of the participants showing more probabilistic behavior, and slightly more than half showing more deterministic behavior. These results corroborate the previously-mentioned results in artificial language learning (Hudson Kam & Newport, 2005; Schumacher et al., 2014; Wonnacott &

Newport, 2005), and extend these findings to a natural-language task in the speaker's L1. As noted, the source of these differences in both the cited articles and the current work is unclear. Nonetheless, this suggests that individual speakers have distinct strategies in language learning and generalization that do not stem from knowledge of existing items in the system, as measured by performance on filler items in these experiments. This indicates that there may be some underlying differences across speakers in how willing they are to generalize lower-frequency patterns to novel forms, versus over-regularizing the most-frequent patterns**.**

In sum, the studies and analyses presented here on these two morphological systems show that speakers of Arabic are able to learn and generalize existing morphological patterns in the language in a manner that reflects the lexical statistics of the system. Moreover, speakers appear to use a combination of coarse- and fine-grained information in drawing analogies to new forms, with the coarse-grained generalization the major basis of similarity. This contrasts with studies of concatenative morphological systems, where fine-grained segmental similarity is the primary basis of similarity in analogy formation (e.g., Alegre & Gordon, 1999; Ernestus & Baayen, 2003). As noted, the evidence for the use of the coarse-grained CV template in generalization is quite strong for the noun plural system, but less clear for the masdar system. Nonetheless, the conjecture that speakers are forming representations on this coarse-grained level for the masdar system is a compelling one, as it would provide a more uniform level of representation for analogy formation and generalization across the verb forms in the verbal paradigm, as well as across different subsystems in Arabic morphology.

Further, these studies shed light on how different types of uncertainty play a role in the decision rule speakers use in deciding amongst candidate forms in linguistic processes. As noted, uncertainty about the optimal morphological pattern to apply to a word can stem from a variety

of sources. In this work, the critical elements of uncertainty are the relative probabilities of the morphological variants, and the number of morphological variants, which can interact in interesting ways. In both the noun plural system and masdar systems, in aggregate speakers match the probabilities of existing morphological patterns even when there are a large number of possible patterns, and the probabilities of these patterns are tracked on an abstract linguistic representation. This contrasts with previous work demonstrating probability-matching when the possible outputs are binary (Ernestus & Baayen, 2003; Hayes et al., 2009; Hudson Kam & Newport, 2005), or when statistics are tracked on the entire system, not on an abstract representation (Hudson Kam, 2009; Walter, 2011). This also contrasts with previous work showing that when there are multiple possible outcomes and the relative probabilities are disparate (e.g., one is much more probable), speakers will over-generalize the more likely outcomes and the lower-probability outcomes will almost or entirely drop out in generalization (Culbertson & Smolensky, 2012; Culbertson et al., 2012; Hudson Kam, 2009). This work demonstrates that speakers can utilize the full range of possible morphological variants in generalization even when some of these possibilities are relatively unlikely, and there are a very large number of possible variants from which to select.

5.2 Implications for future research and future directions

Although this thesis devotes a great deal of attention to how the different aspects of complexity in Arabic morphology make it difficult to learn and generalize, the specific factors that may contribute to this difficulty have not been entirely disentangled. Specifically, I cite two main reasons why both the noun plural and masdar systems should be relatively difficult to learn.

First, the large number of possible patterns, and somewhat irregular correspondence between input and output forms leads to uncertainty about the appropriate output for a particular input form. Second, the non-concatenative morphological patterns require coarse-grained or abstract generalizations across many different word forms in order to adequately capture the CV template. However, it is not clear from this work the extent to which each of these factors contribute to difficulty in learnability, or if they interact in some way. There is some evidence from the developmental literature that Arabic noun plurals are difficult to learn relative to English or German noun plurals (Ravid & Farah, 1999); however, the specific reasons for this difficulty are conflated in the cited work.

As noted previously, there is evidence from experiments using artificial language paradigms that the amount of uncertainty in a system affects learnability. This uncertainty can stem from multiple sources. In particular, in the current work and in the literature there are two main sources of uncertainty: the number of possible outcomes, and the relative probabilities of those outcomes. The majority of the work in this area has focused on relative probabilities without introducing more than two variants, and generally finds that probability-matching versus regularization in adult speakers is at least partially a function of the relative dominance of one variant (Hudson Kam & Newport, 2005; Schumacher et al., 2014). As noted, in artificial language paradigms where there are a large number of outcomes (5+) and one variant is relatively more probable, both children and adults tend toward regularization (Hudson Kam & Newport, 2009). There is still a clear need to further disentangle the relative roles of these two sources of uncertainty in language learning and generalization. In addition, there is evidence both in this thesis and in the artificial language literature (Hudson Kam & Newport, 2005; Schumacher et al., 2014; Wonnacott & Newport, 2005) that individual differences play a role in

the tendency of a speaker to probability-match versus regularize in linguistic generalization. The exact source of these differences across speakers is unclear, and there is still significant work to be done in determining: 1) how these individual differences arise; and 2) how they interact with uncertainty, both in terms of the number of possible outcomes and the relative probabilities of the outcomes in the linguistic system. Further, these types of individual differences have been observed in non-linguistic domains, such as visual categorization (Nosofsky & Johansen, 2000) and probability judgment (Kareev, Lieberman, & Lev, 1997; West & Stanovich, 2003). The fact that we observe these patterns of individual variation in generalization in both linguistic and non-linguistic domains suggests that the source of this variation may reflect some domain-general aspect of information processing. Further research is necessary to examine the extent to which these individual differences in generalization strategy are consistent across domains, and how these differences arise.

With regards to difficulty in learning coarse-grained representations, the evidence is somewhat mixed. There is limited evidence showing that certain types of non-adjacent phonological dependencies are relatively easy to learn. Newport and Aslin (2004) examined statistical learning of non-adjacent dependencies in an artificial language learning task. Adult learners were able to learn non-adjacent segment dependencies when the segments in question were vowels, as well as when the segments were consonants. These conditions, respectively, roughly correspond to learning vowel harmony relations, and the verbal root in Arabic and Hebrew. Adult learners were unable, however, to learn non-adjacent syllable dependencies. Thus, although adult speakers may be able to learn novel generalizations that involve the same segments in non-adjacent position, which is akin to the verbal root in Arabic, learning a representation such as the CV template requires a higher degree of abstraction, as the learner

must generalize across a variety of different segments, where the only regularity is the position and consonantal or vocalic status of the segment. Thus, although there is psycholinguistic evidence that the CV template is active in morphological processing (Boudelaa & Marslen-Wilson, 2004), and it seems clear from the evidence presented in this thesis that adult native speakers use this morphological representation in generalization to unseen forms, there is still a need to experimentally examine the relative difficulty in learning this type of coarse-grained generalization.

Further, as mentioned above, it is unclear to what extent the factors of number of possible variants, relative probability of the variants, and non-concatenativity interact in learnability of a morphological system. A clear next step toward disentangling the effects of these factors in learnability is to directly pit these factors against each other in an artificial language paradigm. In addition, studies on the learning trajectory of Arabic-speaking children's acquisition of the masdar would provide valuable insight into the differences observed in learnability and generalization of these two morphological systems.

References

Abu-Rabia, S. (1998). Reading Arabic texts: Effects of text type, reader type and vowelization. *Reading and Writing: An Interdisciplinary Journal, 10*, 105-119.

Abu-Rabia, S. (2001). The role of vowels in reading Semitic scripts: Data from Arabic and Hebrew. *Reading and Writing: An Interdisciplinary Journal, 14*, 39-59.

Abu-Rabia, S. (2002). Reading in a root-based-morphology language: the case of Arabic. *Journal of Research in Reading, 25*(3), 299-309.

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

Albright, A. (2002). Islands of reliability for regular morphology: Evidence from Italian. *Language, 78*(4), 684-709.

Albright, A. (2009). Feature-based generalisation as a source of gradient acceptability. *Phonology, 26*(01), 9.

Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: a computational/experimental study. *Cognition, 90*(2), 119-161.

Alegre, M., & Gordon, P. (1999). Rule-based versus associative processes in derivational morphology. *Brain and Language, 68*, 347-354.

Attia, M., Pecina, P., Toral, A., Tounsi, L., & van Genabith, J. (2011). A Lexical Database for Modern Standard Arabic Interoperable with a Finite State Morphological Transducer *Systems and Frameworks for Computational Morphology: Proceedings of the 2nd International Workshop, SFCM 2011, Zurich, Switzerland, August 26, 2011* (Vol. 98-118). Berlin: Springer

Baayen, H., & Lieber, R. (1991). Productivity and English derivation: a corpus-based study. *Linguistics, 29*, 801-843.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keeping it maximal. *Journal of Memory and Language, 68*, 255-278.

Becker, M., Ketrez, N., & Nevins, A. (2011). The Surfeit of the Stimulus: Analytic biases filter lexical statistics in Turkish laryngeal alternations. *Language, 87*(1), 84-125.

Berent, I., Marcus, G., Shimron, J., & Gafos, A. (2002). The scope of linguistic generalizations: evidence from Hebrew word formation. *Cognition, 83*, 113-139.

Berko, J. (1958). The child's learning of English morphology. *Word, 14*(150-177).

Berman, R. (1981). Children's regularization of plural forms. *Stanford Papers and Reports on Child Language Development, 20*.

Boudelaa, S., & Gaskell, M. (2002). A re-examination of the default system for Arabic plurals. *Language and Cognitive Processes, 17*(3), 321-343.

Boudelaa, S., & Marslen-Wilson, W. D. (2004). Abstract morphemes and lexical representation: the CV-Skeleton in Arabic. *Cognition, 92*(3), 271-303.

Boudelaa, S., & Marslen-Wilson, W. D. (2010). Aralex: A lexical database for Modern Standard Arabic. *Behavior Research Methods, 42*(2), 481-487.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth.

Broe, M. (1993). *Specification Theory: The treatment of redundancy in generative phonology.* (PhD), University of Edinburgh, Edinburgh.

Brustaad, K., Al-Batal, M., & Al-Tonsi, A. (2004). *Al-Kitaab fii Ta'allum al-'Arabiyya with DVDs: a textbook for beginning Arabic* (Second ed. Vol. I). Washington, D.C.: Georgetown University Press.

Buckwalter, T. (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). Dordrecht: Springer

Buckwalter, T. (2004). Buckwalter Arabic Morphological Analyzer Version 2.0. Philadelphia: Linguistic Data Consortium. Retrieved from http://www.qamus.org

Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes, 10*, 425-455.

Bybee, J., & Moder, C. (1983). Morphological classes as natural categories. *Language, 59*(2), 251-270.

Bybee, J., & Slobin, D. (1982). Rules and schemas in the development and use of the English past tense. *Language, 58*(2), 265-289.

Clahsen, H., Rothweiler, M., Woest, A., & Marcus, G. F. (1992). Regular and Irregular Inflection in the Acquisition of German Noun Plurals. *Cognition, 45*(3), 225-255.

Coleman, J., & Pierrehumbert, J. (1997). *Stochastic phonological grammars and acceptability.* Paper presented at the 3rd Meeting of the ACL Special Interest Group in Computational Phonology, Somerset, NJ.

Cover, T. M., & Thomas, J. A. (1991). *Elements of Information Theory*. New York: Wiley.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS One, 8*(3), e57410.

Culbertson, J., & Smolensky, P. (2012). A Bayesian model of biases in artificial language learning: the case of a word-order universal. *Cognitive Science, 36*(8), 1468-1498.

Culbertson, J., Smolensky, P., & Legendre, G. (2012). Learning biases predict a word order universal. *Cognition, 122*(3), 306-329.

Cutler, A., Sebastian-Galles, N., Soler-Vilageliu, O., & van Ooijen, B. (2000). Constraints of vowels and consonants on lexical selection: Cross-linguistic comparison. *Memory & Cognition, 28*(5), 746-755.

Daelemans, W., Gillis, S., & Durieaux, G. (1994). The acquisition of stress, a data-oriented approach. *Computational Linguistics, 20*, 421-451.

Dawdy-Hesterberg, L., & Pierrehumbert, J. (2014). Learnability and generalization of Arabic broken plural nouns. *Language, Cognition & Neuroscience, 29*(10), 1268-1282.

Dehdari, J. (2009). AraMorph Fast 1.2.1. Retrieved from http://sourceforge.net/projects/aramorph/

Derwing, B., & Skousen, R. (1994). Productivity and the English past tense: Testing Skousen's analogy model. In S. Lima, R. Corrigan & G. Iverson (Eds.), *The reality of linguistic rules: Studies in language companion* (pp. 193-218). Amsterdam: John Benjamins

Deutscher, G. (2001). On the mechanisms of morphological change. *Folia Linguistica Historica, 22*(1-2), 41-48.

Ernestus, M., & Baayen, H. (2003). Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language, 79*(1), 5-38.

Frisch, S., Pierrehumbert, J., & Broe, M. (2004). Similarity avoidance and the OCP. *Natural Language & Linguistic Theory, 22*, 179-228.

Frisch, S., & Zawaydeh, B. (2001). The psychological reality of OCP-Place in Arabic. *Language, 77*(1), 91-106.

Gagliardi, A., Feldman, N., & Lidz, J. (2012). When suboptimal behavior is optimal and why: Modeling the acquisition of noun classes in Tsez. *Proceedings of the 34th Annual Conference of the Cognitive Science Society*.

Gagliardi, A., & Lidz, J. (2014). Statistical insensitivity in the acquisition of Tsez noun classes. *Language, 90*(1), 58-89.

Goldrick, M., & Larson, M. (2008). Phonotactic probability influences speech production. *Cognition, 107*(3), 1155-1164.

Grenat, M. H. (1996). *Argument Structure and the Arabic Masdar.* (PhD), University of Essex.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations, 11*(1), 10-18.

Hammond, M. (1988). Templatic transfer in Arabic broken plurals. *Natural Language & Linguistic Theory, 6*(2), 247-271.

Harrell, R. S. (1962). *A Short Reference Grammar of Moroccan Arabic*. Washington: Georgetown University Press.

Harris, Z. (1954). Distributional Structure. *Word, 10*, 146-162.

Hay, J. B. (2003). *Causes and Consequences of Word Structure*. New York and London: Routledge.

Hay, J. B., & Baayen, R. H. (2005). Shifting paradigms: gradient structure in morphology. *Trends in Cognitive Science, 9*, 342-348.

Hayes, B., Zuraw, K., Siptar, P., & Londe, Z. (2009). Natural and unnatural constraints in Hungarian vowel harmony. *Language, 85*(4), 822-863.

Haykin, S. (1998). *Neural Networks: A comprehensive foundation* (2nd Edn. ed.). Englewood Cliffs: Prentice Hall.

Holes, C. (2004). *Modern Arabic: Structures, functions and varieties*. Washington, D.C.: Georgetown University Press.

Huang, F., Ahuja, A., Downey, D., Yang, Y., Guo, Y., & Yates, A. (2014). Learning Representations for Weakly Supervised Natural Language Processing Tasks. *Computational Linguistics, 40*(1), 85-120.

Hudson Kam, C. L. (2009). More than words: Adults learn probabilities over categories and relationships between them. *Language Learning and Development, 5*(2), 115-145.

Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: the roles of adult and child learners in language formation and change. *Language Learning and Development, 1*(2), 151-195.

Hudson Kam, C. L., & Newport, E. L. (2009). Getting it right by getting it wrong: when learners change languages. *Cognitive Psychology, 59*(1), 30-66.

Kareev, Y., Lieberman, I., & Lev, M. (1997). Through a narrow window: sample size and the perception of correlation. *Journal of Experimental Psychology: General, 126*(3), 278-287.

Kremers, J. (2012). Arabic verbal nouns as phonological head movement. *Working Papers of SFB 732 'Incremental Specification in Context'*.

Krippendorff, K. (1980). *Content Analysis: an Introduction to its Methodology*. Beverly Hills: Sage Publications.

Kruskal, J. B. (1983). An overview of sequence comparison - time warps, string edits, and macromolecules. *Siam Review, 25*(2), 201-237.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics, 22*(1), 79-86.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady, 10*, 707-710.

Levy, M. M. (1971). *The plural of the noun in Modern Standard Arabic.* (PhD), University of Michigan, Ann Arbor.

Lidstone, G. J. (1920). Note on the general case of the Bayes-Laplace formula for inductive or *a posteriori* probabilities. *Transactions of the Faculty of Actuaries, 8*, 182-192.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, 281-297.

Marcus, G., Brinkmann, U., Clahsen, H., Wiese, R., & Pinker, S. (1995). German inflection: The exception that proves the rule. *Cognitive Psychology, 29*, 189-256.

Marcus, G., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., Xu, F., & Clahsen, H. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development, 57*(4), 1-178.

McCarthy, J. (1981). A prosodic theory of nonconcatenative morphology. *Linguistic Inquiry, 12*(3), 373-418.

McCarthy, J. (1982). Prosodic templates, morphemic templates, and morphemic tiers. In H. van der Hulst & N. Smith (Eds.), *The Structure of Phonological Representations* (Vol. I). Dordrecht: Foris

McCarthy, J. (1985). *Formal problems in Semitic phonology and morphology*. New York: Garland Publishing.

McCarthy, J. (1986). OCP effects: gemination and antigemination. *Linguistic Inquiry, 17*(2), 207-263.

McCarthy, J. (1993). *Template form in prosodic morphology*. Paper presented at the Third Annual Formal Linguistics Society of Midamerica Conference, Bloomington.

McCarthy, J., & Prince, A. (1990a). Foot and word in prosodic morphology: The Arabic broken plural. *Natural Language & Linguistic Theory, 8*(2), 209-283.

McCarthy, J., & Prince, A. (1990b). Prosodic morphology and templatic morphology. In M. Eid & J. McCarthy (Eds.), *Perspectives on Arabic Linguistics: Papers from the Second Symposium on Arabic Linguistics*. Amsterdam: J. Benjamins

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *Proceedings of the ICLR Workshop*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing System*, 3111-3119.

Mohammed, E. (2009). *List of Arabic broken plurals*. Retrieved from: http://jones.ling.indiana.edu/~emadnawfal/arabicPlural.txt

Mosteller, F., & Tukey, J. W. (1968). Data analysis, including statistics. In G. Lindzey & E. Aronson (Eds.), *Handbook of Social Psychology* (Vol. 2). Reading, MA: Addison-Wesley

Nakisa, R., Plunkett, K., & Hahn, U. (2001). A cross-linguistic comparison of single and dual-route models of inflectional morphology. In P. Broeder & J. Murre (Eds.), *Models of Language Acquisition: Inductive and deductive approaches*. Cambridge, MA: MIT Press

Newport, E., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology, 48*(2), 127-162.

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

Nosofsky, R., & Johansen, M. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletic & Review, 7*(3), 375-402.

Omar, M. (1973). *The Acquisition of Egyptian Arabic as a Native Language*. The Hague: Mouton.

Parker, R., Graff, D., Chen, K., Kong, J., & Maeda, K. (2011). *Arabic Gigaword Fifth Edition*. Philadelphia: Linguistic Data Consortium.

Pierrehumbert, J. (2001). Why phonological constraints are so coarse-grained. *Language and Cognitive Processes, 16*(5-6), 691-698.

Plunkett, K., & Nakisa, R. (1997). A connectionist model of the Arabic plural system. *Language and Cognitive Processes, 12*(5/6), 807-836.

Prasada, S., & Pinker, S. (1993). Generalisation of regular and irregular morphological patterns. *Language and Cognitive Processes, 8*(1), 1-56.

Prunet, J.-F. (2006). External evidence and the Semitic root. *Morphology, 16*(1), 41-67.

Rabiner, L. R. (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257-286.

Racz, P., Becker, C., Hay, J. B., & Pierrehumbert, J. (2014). 'Rules', 'Analogy' and Social Factors codetermine past-tense formation patterns in English. *Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM* 55-63.

Ratcliffe, R. R. (1998). *The 'Broken' Plural Problem in Arabic and Comparative Semitic: Allomorphy and analogy in non-concatenative morphology*. Amsterdam: John Benjamins.

Ravid, D., & Farah, R. (1999). Learning about noun plurals in early Palestinian Arabic. *First Language, 19*(56), 187-206.

Rennie, J., Shih, L., Teevan, J., & Karger, D. (2003). Tackling the Poor Assumptions of Naive Bayes Text Classifiers. *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 3*, 616-623.

Rosenblatt, F. (1961). *Principles of Neurodynamics: Perceptrons and the theory of brain mechnisms*. Washington, DC: Spartan Books.

Rumelhart, D., & McClelland, J. (1986). On learning the past tenses of English verbs: Implicit rules or parallel distributed processing? In J. McClelland, D. Rumelhart & P. R. Group (Eds.), *Parallel Distributed Processing: Explorations in the microstructure of cognition*. Cambridge, MA: MIT Press

Ryding, K. (2006). *A Reference Grammar of Modern Standard Arabic*. Cambridge: Cambridge University Press.

Scheindlin, R. (2007). *501 Arabic Verbs*. Hauppauge, NY: Barron's Educational Series, Inc.

Schnoebelen, T., & Kuperman, V. (2010). Using Amazon Mechanical Turk for linguistic research. *Psihologija, 43*(4), 441-464.

Schumacher, R. A., Pierrehumbert, J., & LaShell, P. (2014). Reconciling inconsistency in encoded morphological distinctions in an artificial language. *Proceedings of the 36th Meeting of the Cognitive Science Society, Quebec City, Canada*.

Singleton, J. L., & Newport, E. L. (2004). When learners surpass their models: the acquisition of American Sign Language from inconsistent input. *Cognitive Psychology, 49*(4), 370-407.

Skousen, R. (1989). *Analogical Modeling of Language*. Dordrecht: Kluwer.

Skousen, R. (1993). *Analogy and Structure*. Dordrecht: Kluwer.

Snider, N., & Diab, M. (2006). Unsupervised induction of Modern Standard Arabic verb classes using syntactic frames and LSA. *Proceedings of the ACL*.

Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavioral Research Methods, 43*(1), 155-167.

Stemberger, J. P., & MacWhinney, B. (1988). Are inflected forms stored in the lexicon? In M. Hammond & M. Noonan (Eds.), *Theoretical Morphology: Approaches in modern linguistics* (pp. 101-116). Sand Diego, CA: Academic Press

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Royal Statistics Society, B36*, 111-147.

Toutanova, K., Klein, D., Manning, C., & Singer, Y. (2003). *Feature-rich part-of-speech tagging with a cyclic dependency network.* Paper presented at the HLT-NAACL 2003.

Toutanova, K., & Manning, C. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, 63-70.

van Ooijen, B. (1996). Vowel mutability and lexical selection in English: evidence from a word reconstruction task. *Memory & Cognition, 24*(5), 573-583.

Versteegh, K. (1977). *Greek Elements in Linguistic Thinking* (Vol. 7). Leiden: Brill.

Walter, M. (2011). Probability-matching in Arabic and Romance morphology. In E. Broselow & H. Ouali (Eds.), *Perspectives on Arabic Linguistics: Papers from the annual symposia on Arabic Linguistics* (Vol. XXII-XXIII, pp. 203-244). Amsterdam/Philadelphia: John Benjamins

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

West, R., & Stanovich, K. (2003). Is probability matching smart? Associations between probabilistic choices and cognitive ability. *Memory & Cognition, 31*(2), 243-251.

Wonnacott, E., & Newport, E. (2005). Novelty and regularization : the effect of novel instances on rule formation. In A. Brugos, M. R. Clark-Cotton & S. Ha (Eds.), *BUCLD 29: Proceedings of the 29th Annual Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press

Wonnacott, E., Newport, E. L., & Tanenhaus, M. K. (2008). Acquiring and processing verb argument structure: distributional learning in a miniature language. *Cognitive Psychology, 56*(3), 165-209.

Wright, W. (1988). *A Grammar of the Arabic Language* (3rd ed.). Cambridge: Cambridge University Press.

Yang, Y., Yates, A., & Downey, D. (2013). Overcoming the Memory Bottleneck in Distributed Training of Latent Variable Models of Text. *Proceedings of NAACL-HLT 2013*, 579-584.