# A Resource for Arabic Broken Plurals

**Kalina Kostyszyn, Yang Liu, Vasudha Varadarajan, Salam Khalifa, Owen Rambow**
Stony Brook University
Stony Brook, NY, USA
Corresponding author: kalina.kostyszyn@stonybrook.edu

**Abstract**

We present a morphological resource for Arabic broken (irregular) plurals. The resource is organized around inflectional classes (ICs) or paradigms, which combine the singular and plural patterns. We relate the ICs to frequency data from the Arabic Treebank. We use the resource to investigate whether a language learner can predict which paradigm a previously unseen verb falls into.

## 1. Introduction

Many languages have irregular forms in some nominal lexemes for the plural form. For example, in English the plural of *mouse* is *mice*, not **mouses*, and in French, the plural of *cheval* 'horse' is *chevaux*, not **chevals*. We also find irregular plurals in Arabic, both in Modern Standard Arabic (MSA) and the Arabic dialects. Following traditional Arabic grammar terminology, they are called "broken plurals". But in all variants of Arabic, there is a salient fact: there are many more lexemes, and many more occurrences in text, of irregular plurals than in, for example, English or French. This raises several issues.

For morphological theory, we can ask what the set of patterns in broken plurals is, and how we can categorize the existing morphological forms. For second language teaching, we may want a comprehensive resource that students or teachers can consult. For studies of child language acquisition, we can ask how a child can learn such a large set of irregular forms.

There has been a lot of work on Arabic morphological analysis and generation over the years (Al-Sughaiyer and Al-Kharashi, 2004; Habash and Rambow, 2006; Altantawy et al., 2010), with several morphological analyzers and generators available (Buckwalter, 2004; Taji et al., 2018). Furthermore, many descriptive Arabic grammars and textbooks for second language learners have lists of examples of broken plurals (see for example Fischer (1972)). However, to our knowledge, there is no comprehensive lexicographic resource in computer-readable format which lists all *inflectional classes* of broken plural formation along with an extensive list of lexemes for each inflectional class. Here, we define an inflectional class (or, IC for short) as a regular pattern of singular-to-plural inflection; this will be elaborated on in section 2. This paper introduces such a resource, called ABL (for Arabic Broken-plural Lexicon), which we will make freely available. ABL includes frequency statistics from corpora.

The paper is structured as follows. In Section 2, we provide the necessary linguistic background for Arabic broken plurals. In Section 3, we summarizes the existing resources we used to create ABL and the procedure we used to extract the information we needed for ABL. Section 4 summarizes some initial findings we have obtained from inspecting the newly compiled ABL. Finally, Section 5 summarizes the resource we have created.

## 2. Linguistic Background

Arabic, like other Semitic languages, has a complex templatic morphology system. Furthermore, a single lemma inflects to a large number of forms through different combinations of morphological features and cliticization. As a result, Arabic is considered to be morphologically rich. Nominals (nouns and adjectives) in general have smaller paradigms (fewer inflected forms) than verbs; however, nominal inflections for features such as gender and number are not as *regular* as their verbal counterparts. Even though gender and number features in nominals have overt affixational markers, in many cases those features are realized without those markers, either by a change in the template or an arbitrary assignment of the feature such as gender assignment for inanimate nouns.

In Modern Standard Arabic (MSA) and the Arabic dialects, the number feature in nominals has three values: singular, dual, and plural. Singular nominals do not take any affixes, whereas dual nominals take gendered suffixes to express duality and gender. MSA dual affixes also express case. Plurals on the other hand are more complex.

Noun plurals in particular can be expressed by either adding a suffix (hence, a regular plural) or by changing the template of the stem (therefore, an irregular plural). Those two categories of plurals are usually referred to as **sound** plurals – سَالِم *sAlim*[1] – (regular) and **broken** plurals – تَكسِير *taksiyr* – (irregular).

The sound plural can take two classes of suffixes, the feminine plural suffix [-At], and the masculine plural suffixes [-uwn] for the nominative case and (in MSA) [-iyn] for the accusative and genitive cases. The masculine suffix can also surface as [-uw] or [-iy] to ex-

[1] In this paper, we use the Buckwalter transliteration (Buckwalter, 2004) to render Arabic orthography in the Latin alphabet. See https://en.wikipedia.org/wiki/Buckwalter_transliteration for a specification.

| Singular | Plural | Root | Gloss | S. template | Pl. template | IC |
|---|---|---|---|---|---|---|
| mufak∼ir مُفَكِّر | mufak∼iruwn مُفَكِّرُون | f.k.r | thinker | muCaC∼ic | muCaC∼iC+uwn | Reg-m |
| Aikti$Af اِكتِشَاف | Aikti$AfAt اِكتِشَافَات | k.ʃ.f | discovery | AiCtiCAC | AiCtiCAC+At | Reg-f |
| rajul رَجُل | rijAl رِجَال | r.d͡ʒ.l | man | CaCuC | CiCAC | IC12 |
| kitAb كِتَاب | kutub كُتُب | k.t.b. | book | CiCaaC | CuCuC | IC21 |
| safiynap سَفِينَة | sufun سُفُن | s.f.n | ship | CaCiyCap | CuCuC | IC29 |
| mihnap مِهنَة | mihan مِهَن | m.h.n | occupation | CiCCap | CiCaC | IC41 |
| bayt بَيت | buyuwt بُيُوت | b.y.t | house | CaCC | CuCUC | IC44 |
| bayt بَيت | abyAt أَبيَات | b.y.t | verse | CaCC | aCCAC | IC92 |

Table 1: Examples of the Arabic plural system. Except for the root, transcriptions are done using Buckwalter's system; the tilde (∼) represents consonant gemination (shedda). The root is given in IPA. The last column shows the inflectonal class (IC).

press the different state features. Note here that a **suffix class**[2] - a class which take a specific set of suffixes in its inflectional paradigm - is distinct from an **inflectional class** - a class that refers specifically to the singular-plural pattern alternation. Additional vowels may appear after both classes of suffixes but they do not play a role on the plural inflection, therefore it is reasonable to ignore them in contexts similar to this study.

Broken plurals on the other hand do not take suffixes, and instead the template of the singular changes into another templates that expresses plurality. While there is a finite number of templates to express plurality, they are by no means reserved for that purpose. Table 1 showcases the different types of plurals and their respective templates. The combination of the two columns referring to the singular and plural templates thus denotes an **inflectional class**, named in the last column. While the lines for *kitAb* and *safiynap*, for example, share a plural template, their differing singular templates mean they belong to different ICs. Similarly, the two lines for *bayt* share a singular template but have different plural templates, and thus also belong to different ICs. This last example is particularly striking because they don't only share a template, but also a root (b.y.t), but they differ in their plural and their meaning. Other work (Dawdy-Hesterberg and Pierrehumbert, 2014) has investigated learnability of broken plurals at the level of syllable structure; inflectional classes encompass this information because it provides the syllabic template for the singular and plural, including short or long vowels and affixes.

[2]A suffix class can also be called a stem class - either is accurate, since it refers to the group of stems that allowed for a set of suffixes. Here, we generally will use 'suffix class'.

## 3. Resources Used

We mainly used three resources: the data files from Buckwalter's Arabic Morphological Analyzer; the Penn Arabic Treebank; and the Cameltool. We introduce these resources in this section,

### 3.1. Morphological Database

We use the CalimaStar Arabic Morphological Database (Taji et al., 2018) to pick out candidates for broken plurals and analyze the forms. The analyzer database consists of tables of stems, prefixes and suffixes containing the morphological information as shown in tables.

These tables are accompanied by three more tables that restrict the combinations of prefix-stem, stem-suffix and prefix-suffix classes. Each stem in the stem table is attributed with its lexeme, pattern, grammatical number, grammatical gender among other features. A surface form number ("form number") is decided based on the presense of the number suffix that can be attached to the stem, whether to indicate the singular, dual, or plural forms. as opposed to the "number" which denotes the actual grammatical number of the lexeme. Based on this intuition, we mark candidate broken plurals in the context of CalimaStar as the plural form of a lexeme associated with a stem having "form number" as singular but the "number" as plural, since the discrepancy between the two attributes for a single stem is likely an indicator of a broken plural.

### 3.2. Penn Arabic Treebank

We used the Penn Arabic Treebank (Maamouri et al., 2004) to create an inventory of nouns and adjectives (nominals), including the following information: lexeme i.e., its lemma, POS (part-of-speech), and index number. Then we generated frequency counts on the Arabic Treebank (which is lemma-annotated).

### 3.3. CAMeL Tools Morphological Generator

We used the morphological generator module from CAMeL Tools (Obeid et al., 2020) along with the CalimaStar database to generate the full feature array of a given plural which allowed us to verify the type of the plural. The input data are ordered by frequency in Penn Arabic Treebank.

## 4. Findings

We viewed our results through the lens of the Tolerance Principle (Yang, 2005), which states that, for a given morphological rule to be productive (that is, applicable to novel items), it must not surpass a given threshold. Specifically, the formula is as follows:

$$e \leq \theta_N, \text{ where } \theta_N = \frac{N}{lnN}$$

In this formula, $N$ refers to the size of a lexicon, while $e$ is the number of exceptions to the rule in consideration. The Tolerance Principle is built from the idea of a serial search, which orders possible alternative patterns before the 'default' rule which applies if an exception is not triggered. Further investigation has shown that the value of $N$ can refer more specifically to the subsection of the lexicon over which a particular rule is applicable. In the case of German, plurals were a point of contention because the rule used for loanword plurals (suffixation of /-*s*/) was not the most common plural pattern (suffixation of /-*(e)n*/) (Marcus et al., 1995). In a standard serial search, the standard plural pattern would be the 'default' due to the number of items in the lexicon it refers to.

However, this investigation showed that this standard plural would not be applicable to the loanword pattern regardless. In instances where the [-*s*] plural could not be used, it reverted to [-*en*], but otherwise the morphophonological requirements of German dictate that [-*s*] must be used in this instance, despite lower frequency. With this restriction, the rule matches the prediction of the Tolerance Principle. This handling of the feminine plural mirrors our approach to ICs and suffix classes.

### 4.1. Regular and irregular plurals in most frequent nominals

Based on the data from Penn Arabic Treebank, among the most frequent 100 Arabic nominals (n./adj.), 48 words have formal plural forms and 31 of them ($\frac{31}{48}$, 64.58 percent) are regular with the -At and -uw(na) suffixes. In other words, 17 out of 48 words have irregular plural forms. Since 17 is greater than $\frac{48}{ln48}$ (where $ln48$ is 12.4), the regular plural formation rule is not productive among the most frequent 100 Arabic nominals.

Among the most frequent 1000 Arabic nominals (n./adj.), 515 words have formal plural forms and 314 of them (60.97 per cent) are regular. 201 out of the 515 words have irregular plurals, which is greater than $\frac{515}{ln515}$ ( where $ln515$ is 82.5). Therefore, based on the data of the most frequent 1000 Arabic nominals, the regular plural formation rule is not productive either.

### 4.2. Suffix classes

Investigating suffix classes also suggested interesting correlations between suffixes associated with said class and productivity of the broken plural within that class. The Arabic Morphological Analyzer has 32 named suffix classes. Some classes have two variants (e.g. **N0** and **N0_L**). The **_L** ending indicates that the stem begins with *l* ل, which need special orthographic rewrite rules to handle certain prefixes. This treatment however has no effect in the way those stems inflect.

For each class we counted the total number of members and the number of broken plurals only. In Table 2, we see that the plurality of nominals belong to two particular suffix classes, named **N** and **Ndip** (Buckwalter, 2004); these two classes have over 2000 broken items, while the 29 other suffix classes have fewer than 500. Out of the 32 total suffix classes, only 8 have more than 100 broken plurals belonging to that class. Because of this, we used a ratio of broken-to-sound plurals within each class, rather than a raw count; these results can be seen in Table 2. While the suffix class **N** has the most members overall, the suffix classes with the highest ratio (and, by extension, highest likelihood of a member being broken) were **Ndip** and **Ndip_L**. The class **Ndip** represent *diptotes*, which are a special class of nominals that have only two distinct inflections for the three cases, where both the accusative and genitive cases inflect the same way. Most of the stems belonging to **Ndip** are a class of broken plurals called *plural of plurals*, which take a specific pattern that is always treated as a diptote.

These suffix classes, of course, are determined by the number and type of suffixes the members of that class are able to take. This ranged from a class only taking a single suffix (named **N0**) to one with 35 possible suffixes (named **Nall**, because it takes all possible suffixes). In this corpus, 'Suff-0' refers to an empty suffix (no change to the stem), but because the suffix class **N0** is entirely defined by members that take no suffixes, we did not remove empty suffixes from consideration.

Using the broken-to-sound ratio and the number of suffix associated with a suffix class, we ran correlations in R using Pearson's $r$ and Kendall's $\tau$. Pearson's $r$ will measure the strength of a correlation between two values, while Kendall's $\tau$ will determine the strength of a ranking of these values. Additionally, it was noted that some of the suffix classes referred to dual nouns; we included this as a binary value referring to whether or

Table 2: General information about the suffix classes. 'Broken' refer to the raw counts of broken plurals in a suffix class, and 'ratio' refers to the broken-to-sound plural ratio within a class.

| | **N** | **Ndip** | **Nap** | **N0_Nh** | **Nh** | **Nhy** | **N0** | **NK** |
|---|---|---|---|---|---|---|---|---|
| Total | 6428 | 2177 | 2862 | 1168 | 769 | 1163 | 892 | 443 |
| Broken | 2265 | 2103 | 391 | 475 | 266 | 354 | 134 | 157 |
| Ratio | 0.55843 | 33.38095 | 0.15914 | 0.70790 | 0.53521 | 0.4425 | 0.19198 | 0.56071 |
| | **Nuwn_Niyn** | **NapAt** | **NAt** | **N/At** | **Nel** | **N_L** | **Nall** | **FW-WaBi** |
| Total | 248 | 1507 | b1323 | 1302 | 396 | 179 | 4852 | 36 |
| Broken | 5 | 5 | 19 | 17 | 60 | 44 | 4 | 7 |
| Ratio | 0.02083 | 0.00334 | 0.01464 | 0.01324 | 0.20761 | 0.32836 | 0.00201 | 0.24138 |
| | **NduAt** | **N0_Nhy** | **Ndu** | **Numb** | **FW-Wa-y** | **FW-Wa** | **FW-Wa-a** | **Napdu** |
| Total | 1188 | 71 | 437 | 8 | 3 | 44 | 40 | 869 |
| Broken | 12 | 51 | 3 | 4 | 2 | 1 | 1 | 8 |
| Ratio | 0.01020 | 2.55 | 0.00694 | 1.0 | 2.0 | 0.02632 | 0.11111 | 0.00935 |
| | **Ndip_L** | **N0_Nh_L** | **Nap_L** | **N0_L** | **Nhy_L** | **NAn_Nayn** | **Nh_L** | |
| Total | 31 | 8 | 109 | 33 | 29 | 693 | 19 | |
| Broken | 27 | 8 | 1 | 6 | 7 | 1 | 6 | |
| Ratio | 27.0 | 0.53333 | 0.00926 | 0.22222 | 0.31818 | 0.00146 | 0.46154 | |

not the dual occurred.

Table 3: Evaluation scores for correlation models.

| **Evaluation** | **Count → ratio** | **Dual → ratio** |
|---|---|---|
| Pearson $r$ | -0.05856712 | -0.1663988 |
| $r$ significance | p=0.7543 | p=0.371 |
| Kendall $\tau$ | -0.2632774 | -0.5295178 |
| $\tau$ significance | p=0.04329 | p=0.000472 |

First, we considered the relation between the broken-to-sound ratio and the suffix counts, and the broken-to-sound ratio and the presence of dual. Results of these correlations can be seen in table 3. Neither of the Pearson models gave significant values, but both of the Kendall values were significant, so we continued with these. With that in mind, both the number of suffixes per suffix class and the presence of dual were negative predictors of a member of a suffix class being broken or not - more suffixes or presence of dual made it less likely for broken plurals.

With this in mind, we created a model that, using both of these parameters, predicted the likelihood of a member of a given suffix class having a sound plural. We made two adjustments, in that the number of suffixes were normalized to values between 0 and 1, and that suffix classes specific to function words (**FW-Wa**, **FW-Wa-a**, and **FW-Wa-y**) were removed from consideration. This gave us the following, where $b$ is the likelihood of a broken plural, $c$ is the count, and $d$ is the presence of dual:

$$b = 2.529 + (3.468 * c) - (4.065 * d)$$

Results of this model would provide a (positive) number that, the higher it is, the less likely for a class to have broken members. For example, for the suffix class **Nall**, which has 35 suffixes and has dual, the value comes out to 119.844 (the highest value), making it very unlikely to have broken plurals. On the other hand, **Ndip** was had the highest broken-to-sound ratio, 6 suffixes, and no dual; this gave it a score of 23.337. Full results of this model are presented alongside our tool.

The mean score was approximately 28.448, of which 18 suffix classes fell below this threshold. While it did properly select the items with broken-to-sound ratios above 1 (per table 2, this includes **Ndip**, **Nip_L**, **N0_Nhy**), it also chose several with very low ratios. Our model predicted that **Nh** and **N0** (as well as **Nh_L** and **N0_L**) would be the 'most' broken, scoring them each 5.997. All of these classes only had a single associated suffix, but also poor broken-to-sound ratios. So, while suffix count and dual are predictors of 'brokenness', there are still outlying features that must be considered.

With regards to the Tolerance Principle, raw counts are insufficient to determine productivity - classes **Ndip** and **N0_Nhy** fail in this regard, despite our model selecting them as candidates for broken plurals. The Tolerance Principle would also reject our model's top candidates **Nh** and **N0** (though, **Nh_L** and **N0_L** are acceptable). At this point, the only possible claim that could successfully invoke the Tolerance Principle would be using the *number of suffix classes* rather than individual items. In this regard, for 28 suffix classes (excluding function words) and only 4 suffix classes that have a broken-to-sound ratio over 1, $4 < \frac{24}{ln24} = 8.4$. So, while we have a model that predicts what suffix classes broken plurals belong to, the Tolerance Principle would then assert that most of these cases are not robust enough to be productive.

## 5. Resource Created

Inflection classes are templates of singular pattern-plural pattern inflections. From the lexical database, we extracted the singular-plural pattern pairs for candidate broken plural lexemes, and we found a total of 1970 inflection classes.

Using the Penn Arabic Treebank (Maamouri et al., 2004), we found overall 1266 Inflectional Classes (ICs) attested among the lexemes occurring in it. The table below displays how many ICs are represented in the most frequent N lexemes in the PATB. For instance, 39 inflectional classes are represented in the most frequent 100 words and 39 out of the 39 ones are among the most common 200 ICs. The "most common ICs" is measured using the Treebank corpus (the sum of frequencies of words with the same IC).

Table 4: Number of ICs found among the most frequent N lexemes

| N | ICs |
|---|---|
| 100 | 39 |
| 200 | 72 |
| 500 | 134 |
| 1000 | 218 |
| 5000 | 504 |
| 10000 | 753 |
| 12222 | 843 |

We will deliver a resource which includes the following components:

1. The **IC Table**: a list of 1,266 inflectional classes (ICs) for MSA nominals, defined as pairs of (singular, plural) patterns for broken plurals, plus the ICs "regular-masc" and "regular-fem".

2. The **IC-to-Lemma Table**: for each IC, a list of nominals lexemes which are part of that IC; these will be presented as sequences of radicals, which together with the singular pattern of the IC will provide the lemma (i.e., singular nominative form).

3. The **Lemma-to-IC Table**: For each nominal lemma, its IC. We will also provide an English gloss for each lexeme.

4. Frequency counts against the Penn Arabic Treebank for each IC (recorded in the IC-to-Lemma Table) and for each lexeme (recorded in the Lemma-to-IC Table). (The IC frequency count is simply the sum of the frequency counts of its lemmas.)

This resource, called ABL, will be distributed as text files, and will be available freely for download.

## 6. Conclusion

In this paper, we presented a new resource: ABL, a computer-readable list of MSA nominal inflectional classes, defined by singular-plural patterns, along with a list of lemmas for that class and frequency counts for each lemma in the Penn Arabic Treebank.

We then used this resource to investigate the question of Arabic broken plurals from the perspective of the Tolerance Principle. First, we look at the regularity of the most frequent 100 and 1000 nominals in Arabic, and find that in both case, they do not meet the threshold for productivity. Then, we examine the suffix classes of Arabic as predictors of broken plurals and how this may correlate with predictions of the Tolerance Principle. We developed a model that, using number of suffixes and presence of the dual, could select the suffix classes that have the highest ratios of broken plurals to sound plural. We used this model to find which suffix classes are most likely to have broken plurals based on these factors and compared them against whether, by the Tolerance Principle's measure, the broken plural would be productive in this class. Using this measure, the 'most broken' suffix class per ratio alone was judged nonproductive, as were two of the four classes predicted 'most broken' by our model. Others can continue to use this information to pursue this vein of research, or any other research based on inflection of Arabic nominals.

## 7. Bibliographical References


Al-Sughaiyer, I. A. and Al-Kharashi, I. A. (2004). Arabic morphological analysis techniques: A comprehensive survey. *Journal of the American Society for Information Science and Technology*, 55(3):189–213.

Altantawy, M., Habash, N., Rambow, O., and Saleh, I. (2010). Morphological Analysis and Generation of Arabic Nouns: A Morphemic Functional Approach. In *Proceedings of the seventh International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta.

Buckwalter, T. (2004). Buckwalter Arabic morphological analyzer version 2.0. Linguistic Data Consortium (LDC) catalog number LDC2004L02, ISBN 1-58563-324-0.

Dawdy-Hesterberg, L. G. and Pierrehumbert, J. B. (2014). Learnability and Generalisation of Arabic Broken Plural Nouns. *Language, Cognition and Neuroscience*.

Fischer, W. (1972). *Grammatik des klassischen Arabisch*. Harrassowitz.

Habash, N. and Rambow, O. (2006). Magead: A morphological analyzer for Arabic and its dialects. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (Coling-ACL'06)*, Sydney, Australia.

Maamouri, M., Bies, A., Buckwalter, T., and Mekki, W. (2004). The penn arabic treebank: Building a

large-scale annotated arabic corpus. In *NEMLAR conference on Arabic language resources and tools*, volume 27, pages 466–467. Cairo.

Marcus, G. F., Brinkmann, U., Clahsen, H., Wiese, R., and Pinker, S. (1995). German inflection: The exception that proves the rule. *Cognitive psychology*, 29(3):189–256.

Obeid, O., Zalmout, N., Khalifa, S., Taji, D., Oudah, M., Alhafni, B., Inoue, G., Eryani, F., Erdmann, A., and Habash, N. (2020). CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France, May. European Language Resources Association.

Taji, D., Khalifa, S., Obeid, O., Eryani, F., and Habash, N. (2018). An Arabic morphological analyzer and generator with copious features. In *Proceedings of SIGMORPHON*, October.

Yang, C. (2005). On Productivity. *Linguistic Variation Yearbook*, 5(1):265–302.