



Calculating Name Frequency in Arabic

by Jack Halpern
December 13, 2009

1. Introduction

Augmenting DANA with frequency data from relevant lexical resources should increase the effectiveness with which it can be used to distinguish names from non-names. By including relevant frequency data, DANA can be used to determine the likelihood of an arbitrary string of Arabic being a name. Because such a system would be purely quantitative and deterministic, it could be easily used for automatically categorizing names and non-names.

2. Examples

Table 1 shows a selection of names taken from CJKI's v1.0 release of DANA. Shaded columns show frequency data taken from various lexical resources, but not yet available in DANA. The inclusion of such frequency information would allow for the creation of a finely-grained **Name Probability** field, which in turn could be used to automatically categorize names and non-names.

For illustrative purposes, in Table 1 variant forms of each Arabic name have been collapsed into a single a single lexeme, the lemma form of which is shown in **Arabic**. This is followed by Table 2 that shows the variants as well.

Table 1. Database of Arabic Names in Arabic (DANA)

Variants Collapsed

Corpus Name Count represents the occurrences of **Arabic** as a name in a tagged corpus.

Corpus Nonname Count represents the occurrences of **Arabic** as a nonname in a tagged corpus.

Name Probability represents the probability of **Arabic** being a name

Phonebook Count represents the occurrences of **Arabic** in CJKI phonebook databases

<i>Intelligence Community Romanization</i>	<i>Arabic</i>	<i>Name Probability</i>	<i>Corpus Name Count</i>	<i>Corpus Nonname Count</i>	<i>Phonebook Count</i>
Hasan	حسن	L4	1 726	4 853	212 429
'Abd-al-'Aziz	عبد العزيز	L0	13	0	120 349
'Ali	علي	L4	52 114	175 912	337 551
al-Misri	المصري	L3	2307	2781	2683
Wafi	وفي	L5	0	23	7
'Abd-al-Shafi	عبد الشافي	L0	-	-	4440
Walis	وليس	L5	-	-	2



Table 1 includes the following frequency data:

- **Corpus Name Count** represents the occurrences of **Arabic** as a name in CJKI's Arabic tagged corpus.
- **Corpus Nonname Count** represents the occurrences of **Arabic** as a nonname in CJKI's Arabic tagged corpus.
- **Phonebook Count** represents the occurrences of **Arabic** in CJKI phonebook databases. CJKI phonebook databases are built using data from Egypt, Saudi Arabia, Jordan, Tunisia, Lebanon, Bahrain, Oman, the Palestinian Territories, the United Arab Emirates, and other countries and jurisdictions.

Table 2. Database of Arabic Names in Arabic (DANA)

Variants Shown

<i>Intelligence Community Romanization</i>	<i>Normalized Arabic</i>	<i>Variant Arabic</i>	<i>Name Probability</i>	<i>Corpus Name Count</i>	<i>Corpus Nonname Count</i>	<i>Phonebook Count</i>
Hasan	حسن	حسن	L4	1726	4853	212 429
'Abd-al-'Aziz	عبد العزيز	عبدالعزیز	L0	13	0	118 158
'Abd-al-'Aziz	عبد العزيز	عبد العزیز	L0	-	-	2191
'Ali	علي	على	L4	49285	164818	274 868
'Ali	علي	علي	L4	2829	11094	78 304
'Ali	علي	عليّ	L4	-	-	-
'Ali	علي	عليّ	L4	-	-	-
al-Misri	المصري	المصري	L3	2307	2781	306
al-Misri	المصري	المصرى	L3	-	-	2377
al-Misri	المصري	المصريّ	L3	-	-	-
al-Misri	المصري	المصرىّ	L3	-	-	-
Wafi	وفي	وفى	L5	0	23	6
Wafi	وفي	وفى	L5	-	-	1
Wafi	وفي	وفىّ	L5	-	-	-
Wafi	وفي	وفىّ	L5	-	-	-
'Abd-al-Shafi	عبد الشافي	عبد الشافي	L0	-	-	4440
'Abd-al-Shafi	عبد الشافي	عبدالشافي	L0	-	-	4440
'Abd-al-Shafi	عبد الشافي	عبد الشافي	L0	-	-	4440



'Abd-al-Shafi	عبد الشافي	عبدالشافى	L0	-	-	4440
Walis	وليس	وليس	L5	-	-	2

3. Calculating Name Probability

The **Name Probability** field is a convenient way of encapsulating all the information contained in **Corpus Name Count**, **Corpus Nonname Count** and **Phonebook Count** fields. It shows the likelihood of a given DANA entry being a name, as opposed to being a non-name. **Name Probability** can take any of the following values:

- [L0] Almost always a name (90%+)
- [L1] Commonly a name (70-90%)
- [L2] Sometimes a name (50%+)
- [L3] Seldom a name (30-50%)
- [L4] Uncommonly a name (10-30%)
- [L5] Almost never a name (-10%)

Where available, tagged corpus data as found in **Corpus Name Count** and **Corpus Nonname Count** fields is probably the best way to calculate **Name Probability**. Simply dividing **Corpus Name Count** by the sum of itself and **Corpus Nonname Count** is sufficient to assign a **Name Probability**. For example, corpus data collectively shows that عبد العزيز is always a name. On the other hand, corpus data shows that although حسن is common as a name, it is even more common as a non-name, so that it needs to be handled with a great deal of caution.

If a DANA entry is not present in CJKI's Arabic tagged corpus, **Phonebook Count** can be used to guess a **Name Probability** value. For example, وليس has a low **Phonebook Count**, which suggests that it is rarely used as a name. In contrast, عبد الشافي has a relatively high **Phonebook Count**. Although the absence of these strings in the CJKI tagged corpus precludes the creation of any definitive **Name Probability** values, assigning L5 to وليس and L0 to عبد الشافي would seem to be reasonable. Of course, more corpus data would improve the accuracy of **Name Probability**. Much larger tagged corpora can be obtained, and even (easily created) untagged corpora of natural written Arabic can be used to greatly improve the accuracy of results.

Using the two **Corpus Counts** in combination with the **Phonebook Count** can be further augmented by making use of **A-FULEX**, probably the world's first **Arabic Full Form Dictionary**, currently under development at CJKI. This is a comprehensive lexicon that contains every conjugated, inflected and cliticized form in the language, which could be used to determine the status of a term with greater precision.